

Precision-based Item Selection for Exposure Control in Computerized Adaptive Testing

By

© 2017

Ian A. Carroll

Submitted to the graduate degree program in the Department of Psychology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairperson: Holger Brandt, Ph.D.

John Poggio, Ph.D.

William Skorupski, Ed.D.

Amber Watts, Ph.D.

Nancy Ann Hamilton, Ph.D.

Date Defended: December 8, 2017

The Dissertation Committee for Ian A. Carroll
certifies that this is the approved version of the following dissertation:

Precision-based Item Selection for Exposure Control in Computerized Adaptive Testing

Chairperson: Holger Brandt, Ph.D.

Date approved: December 8, 2017

Abstract

Item exposure control is, relative to adaptive testing, a nascent concept that has emerged only in the last two to three decades on an academic basis as a practical issue in high-stakes computerized adaptive tests. This study aims to implement a new strategy in item exposure control by incorporating the standard error of the ability estimate into the item selection process. A new method, which is a simple modification of an existing and widely implemented method, is evaluated with respect to quality of ability estimation, control of item exposure, and vulnerability to actions to compromise the validity of a test. The new method is compared to existing methods using statistical simulation. Results suggest that the new method performs adequately in all outcomes and across simulation conditions. The study concludes with a discussion of the potential implications of the findings, as well as promising future avenues of research.

Acknowledgements

I dedicate this dissertation to my deceased mother, Eugenia. Robbed by the fickle and seemingly random nature of fate, we were unable to share some of the moments that have thus far defined my adult life: my marriage to Melody, the birth of our son, Marcus, and the completion of my Master's and Doctoral degrees. I hope you have found some measure of peace in the afterlife, and that I have made you proud.

I would also like to mention my father, John, who in my youth toiled endlessly to provide me and my siblings with the unique opportunities we enjoyed, and who over time has transitioned into becoming a mentor and friend. In this document I commit to written record the fact that your selflessness in caring for and providing for your children has not been in vain, and that your unambiguously positive effect upon the world already reverberates not only through your patients but also your children and their children.

To my wife, Melody, I say that I love you and that I have never been so happy as this period in my life where I finally met you. This dissertation is the culmination of long and arduous processes, academic and real-world, in which you were directly involved and on which you selflessly applied your myriad skills in order to compensate for obvious deficiencies in my own. I look forward to our future together, standing here at the beginning of this long adventure, and would rather have no one but you by my side. Also, to my son, Marcus: you cannot yet read, or write, or even speak fluently, but all those things will come in time. I have come to know you in these two wonderful years that your mother and I have had you on this Earth, and I consider it a great privilege and responsibility to be your father. It is a blessing to have you here on this journey with us.

I thank Melody's parents, Jau-Jiun and Tzu-Jung, for their ongoing support of our family and for the existence of my wonderful wife in the first place. I hope and suspect we soon will have an opportunity to once again come together and be happy in each-other's company.

In my family I additionally will single out my sister Grace; we have helped each-other during good and bad times, some of which occurred over the long duration this study was conducted, and I hope this continues in perpetuum. It has been a source of happiness for me seeing your life blossom into the great thing that it inevitably will be.

I thank many Professors, both past and present, from the University of Kansas who have helped me enormously in becoming who I am today. My committee for this dissertation consisted of Dr. Holger Brandt, Dr. Amber Watts, Dr. Nancy Hamilton, Dr. John Poggio, and Dr. William Skorupski. I thank each of you here for being willing to be part of this project and for all the help you provided along the way.

Carol Woods, my first advisor, from whom I learned much of what I know about factor analysis, differential item functioning, psychometrics in general, and the true path one must take in order to succeed -- the one where, beset on all sides by failure and criticism, you persist, learn, and ultimately achieve your goals.

Wei Wu, my second advisor, advised me not only on this project, but also taught me much of what I know about missing data and complex topics pertaining to SEM. Thank you for being so understanding and kind to me, especially when I was at my most vulnerable, and for providing me with so much support not only in my teaching but also in finding and conducting collaborative work.

Pascal DeBoeck taught me across many different classes like nonparametric, longitudinal, and categorical. I benefited greatly not only because of the knowledge but also because of the unusually cogent manner in which it was conveyed.

William Skorupski taught me much of what I know about item response theory, bayesian statistics, and programming in general. He also provided mentoring during some difficult periods of graduate school, the value of which was enormous both in its content and its timely nature. I remember all of it: thank you for everything, Billy.

John Poggio was Melody's advisor, a speaker at our wedding, a member of this dissertation committee, and a font of knowledge for presumably everyone he comes into contact with. Some of the truly fundamental questions addressed in this dissertation arose out of discussions we had regarding the topic -- and insofar as knowing how to do statistics and knowing how to think are entirely different things, I am grateful to have been a part of those conversations.

Holger Brandt was the chair of this dissertation project and an excellent person to bounce ideas off of in regard to essentially anything: statistical considerations, methodological considerations, writing, presenting images, presenting ideas... you name it. I had never thought I would get so much useful guidance from an advisor who entered the program after I finished my coursework, but naturally there is always more to learn and I am thankful to Holger for whatever parts of this manuscript the current reader perceives to be particularly well-articulated or presented.

I will also list fellow graduate students from whom I learned and with whom I commiserated over the many years it has taken me to get to this point: Jared Harpole, Mian Wang, Graham Rifkenbark, Po-Yi Chen, Richard Kinai, Benjamin Kite, Leslie Shaw, Fan Jia,

Terrence Jorgensen, Elizabeth Grandfield, Mauricio Guarnier-Villareal, Luke McCune, Kyle Lang, and presumably many others. Thank you for the camaraderie.

Table of Contents

Abstract	iii
Acknowledgements	iv
Table of Contents	viii
List of Tables	x
List of Equations	xi
List of Figures	xii
Chapter I: Introduction.....	1
A Background on Item Exposure	1
Statement of Problem	4
Chapter II: Literature Review	6
Initial Development of Item Exposure Control Methods	6
Multi-Stage Tests and Item Exposure.....	9
Item Influence	18
Item Bank Management.....	24
Chapter III: Methodology	27
Proposed Method	27
Proposed Method: Caveats and Justifications	32
Software	33
Pilot Study	33
Primary Study	35
Pilot Study Results.....	37
Chapter IV: Results.....	43

Estimation of Ability and Standard Error of the Ability Measure.....	43
Section 1: Bias.	43
Section 2: Standard Error of the Ability Measure	45
Item Exposure.....	73
Section 1: Exposure in aggregate.....	73
Section 2: Exposure in relation to item properties.....	77
Section 3: Vulnerability to item pilfering.	91
Chapter V: Discussion	98
Part 1: General Observations	98
Part 2: Limitations	103
Part 3: Considerations for Implementation in Existing Testing Regimen	104
Part 4: Future Research Directions	106
References.....	108

List of Tables

Table 1. Item parameters in the pilot simulation	34
Table 2. Other parameters included in the simulation	34
Table 3. Information on simulation conditions	36
Table 4. Information on to-be-evaluated outcome measures	36
Table 5. Modified parameters in the item generation process	36
Table 6. Summary statistics of final $SE(\theta)$ estimates by method ($k=500$)	58
Table 7. Summary statistics of final $SE(\theta)$ estimates by method ($k=1,000$)	59
Table 8. Summary statistics of final $SE(\theta)$ estimates by method ($k=1,500$)	59

List of Equations

1. Linear integer formula for maximum information.....	8
2. uMFS: Average error variance in a θ estimate	16
3. uMFS: Variability in the precision of the θ estimate for all values of θ	16
4. uMFS: Testlet information function for the 1st test stage	16
5. uMFS: Proportion of examinees routed to each subsequent testlet	17
6. Weighted linear combination of multiple operational criteria	17

List of Figures

Figure 1. A three-state multistage testing design.....	11
Figure 2. Item exposure control test administration method	14
Figure 3. On-the-Fly assembled multistage adaptive testing.....	23
Figure 4. Linear program of maximum item exposure constraints.....	28
Figure 5. Additional functions necessary for the calculations in Figure 4.	29
Figure 6. $SE(\theta)$ trajectory by method in pilot study results ($k=1,000$)	37
Figure 7. $SE(\theta)$ trajectory by method in pilot study results ($k=10,000$)	38
Figure 8. Comparison on item exposure control by method.....	39
Figure 9. Comparison on item exposure control by method.....	40
Figure 10. Percentage overlap in item selection by method	42
Figure 11. Empirical densities of true and estimated ability ($k=500$).....	44
Figure 12. Empirical densities of true and estimated ability ($k=1,000$).....	44
Figure 13. Empirical densities of true and estimated ability ($k=1,500$).....	45
Figure 14. Comparison of final estimates for standard error of the ability estimates within UE-ST Conditions	46
Figure 15. Comparison of final estimates for standard error of the ability estimates within PB-ST conditionss	47
Figure 16. Comparison of final estimates for standard error of the ability estimate within CAT conditions.....	48
Figure 17. Comparison of final estimates for standard error of the ability estimate within Nonadaptive conditions	49
Figure 18. Averages of $SE(\theta)$ trajectory by method in simulated examination ($k=500$).....	50

Figure 19. Averages of SE (θ) trajectory by method in simulated examination ($k=1,000$).....	51
Figure 20. Averages of SE (θ) trajectory by method in simulated examination ($k=1,500$).....	52
Figure 21. Relationship between true ability and standard error estimates within UE-ST conditions.....	54
Figure 22. Relationship between true ability and standard error estimates within CAT conditions	55
Figure 23. Relationship between true ability and standard error estimates within PB-ST conditions.....	56
Figure 24. Relationship between true ability and standard error estimates within Nonadaptive conditions.....	57
Figure 25. Box-plots of SE(θ) within UE-ST conditions.....	60
Figure 26. Box-plots of SE(θ) within CAT conditions.....	62
Figure 27. Box-plots of SE(θ) within Nonadaptive conditions	63
Figure 28. Box-plots of SE(θ) within PB-ST conditions.....	64
Figure 29. Q-Q plots of SE(θ) estimate distributions: CAT vs PB-ST ($k=500$).....	66
Figure 30. Q-Q plots of SE(θ) estimate distributions: CAT vs PB-ST ($k=1,000$).....	67
Figure 31. Q-Q plots of SE(θ) estimate distributions: CAT vs PB-ST ($k=1,500$).....	67
Figure 32. Q-Q plots of SE(θ) estimate distributions: Nonadaptive vs PB-ST ($k=500$).....	69
Figure 33. Q-Q plots of SE(θ) estimate distributions: Nonadaptive vs PB-ST ($k=1,000$).....	69
Figure 34. Q-Q plots of SE(θ) estimate distributions: Nonadaptive vs PB-ST ($k=1,500$).....	70
Figure 35. Q-Q plots of SE(θ) estimate distributions: UE-ST vs PB-ST ($k=500$).....	71
Figure 36. Q-Q plots of SE(θ) estimate distributions: UE-ST vs PB-ST ($k=1,000$).....	72
Figure 37. Q-Q plots of SE(θ) estimate distributions: UE-ST vs PB-ST ($k=1,500$).....	72

Figure 38. Rank order of frequency of item usage within CAT conditions	74
Figure 39. Rank order of frequency of item usage within PB-ST conditions.....	75
Figure 40. Rank order of frequency of item usage within UE-ST conditions	76
Figure 41. Rank order of frequency of item usage within Nonadaptive conditions	77
Figure 42. Discrimination parameter and frequency of item usage within UE-ST conditions.....	79
Figure 43. Discrimination parameter and frequency of item usage within PB-ST conditions	80
Figure 44. Discrimination parameter and frequency of item usage within CAT conditions	81
Figure 45. Discrimination parameter and frequency of item usage within Nonadaptive conditions	82
Figure 46. Difficulty parameter and frequency of item usage within UE-ST conditions.....	84
Figure 47. Difficulty parameter and frequency of item usage within PB-ST conditions	85
Figure 48. Difficulty parameter and frequency of item usage within CAT conditions	86
Figure 49. Difficulty parameter and frequency of item usage within Nonadaptive conditions....	87
Figure 50. Guessing parameter and frequency of item usage within UE-ST conditions.....	88
Figure 51. Guessing parameter and frequency of item usage within PB-ST conditions	89
Figure 52. Guessing parameter and frequency of item usage within CAT conditions	90
Figure 53. Guessing parameter and frequency of item usage within Nonadaptive conditions.....	91
Figure 54. Pairwise comparison on shared items by score on the latent trait ($k=500$)	92
Figure 55. Pairwise comparison on shared items by score on the latent trait ($k=1,000$)	93
Figure 56. Pairwise comparison on shared items by score on the latent trait ($k=1,500$)	94
Figure 57. Overlap percentage amongst high ability subjects by method ($k=500$)	95
Figure 58. Overlap percentage amongst high ability subjects by method ($k=1,000$)	96
Figure 59. Overlap percentage amongst high ability subjects by method ($k=1,500$)	96

Figure 60. Rank order by item selection frequency ($k=500$)	99
Figure 61. Rank order by item selection frequency ($k=1,000$)	100
Figure 62. Rank order by item selection frequency ($k=1,500$)	101

Chapter I: Introduction

A Background on Item Exposure

Item exposure is a crucial concept in the administration of high-stakes tests. In simplest terms, it is how often an item is used during a test. Some typical qualifiers for “exposure” are “over” and “under”, with over meaning that an item was used too many times according to some criterion, and under meaning that an item was not used enough.

An item is considered to be exposed in each instance it is administered to an examinee. Exposure refers to the frequency, especially in relation to other items that could have been administer, that one item was administered. In most testing environments, having items that are overexposed relative to others is not an issue: a physician asking a patient if he/she has had chest pain, for example, is a perfectly ordinary occurrence and is understandably more likely than more condition-specific questions (e.g. “How have you been faring in the wake of your recent trans-catheter aortic valve replacement?”). One might say here that the assessment process in this example does not suffer from this overexposure, because there is no correct answer to the chest pain question. In such an environment, subjects can know many of the questions ahead of time (and, indeed, should) without any threat to the validity of the assessment process.

Therefore, item exposure is perhaps an issue that is idiosyncratic to high-stakes testing, in the sense that it is only problematic if an item is repeatedly given when the item has a legitimate “correct” answer. The distinction here is sufficiently obvious that it is essentially axiomatic: to borrow from a previous example, it is perfectly acceptable to know many of the questions a physician might ask, but it is probably not good if students know many of the questions on an upcoming math test. As such, the problems described in the item exposure literature almost invariably feature the GRE, SAT, MCAT, and other high-stakes tests.

One might view the issue of item exposure as an economic one; testing companies do not have access to infinite resources, so they can only produce and vet a finite number of items with which they design and administer tests. This conundrum is perhaps best explained in “Rescuing Computerized Testing by Breaking Zipf’s Law” (Wainer, 2000), which contains a wealth of information relating to the cause, costs, and potential countermeasures for the “item exposure” problem.

As described in the article, Educational Testing Service (ETS) sued Kaplan in 1995 over “item theft”. Essentially what this amounted to is Kaplan sending twenty to thirty “thieves” to take the GRE, with their sole purpose being to remember and subsequently write down difficult test questions. Within an alarmingly short period of time (allegedly less than a month), Kaplan could successfully reproduce approximately 80% of the difficult items from ETS’ item bank for that subsection of the CAT GRE.

This is a direct threat to the test’s validity (where we might operationalize validity as a scenario in which responses to questions on the test are determined solely by the characteristics the test ostensibly is designed to evaluate), and one that simply cannot go unacknowledged. The issue that Dr. Wainer expounds upon in the article is that the more obvious countermeasures will be completely ineffective due to Zipf’s law.

Zipf’s law, named after its presumed discoverer Dr. George K. Zipf in the mid 20th century, is a phenomenon wherein the rank order of event frequency is linearly related to the log of the events’ frequencies. This is shown throughout Dr. Wainer’s article in various contexts, and the consistency of the effect is rather remarkable, which explains why it is called a “law” despite no mathematical proof establishing the truth of the assertion (one might say the “law” has been established on an empirical basis).

This relates to computerized adaptive testing because Zipf's law also applies to the frequency of item selection; as demonstrated graphically in the article, there exists a relatively linear relationship between the log of how often an item was selected versus the rank order of the frequently selected items.

This means that increasing the size of the item bank is a highly inefficient strategy to combat practices of companies similar to Kaplan. As stated by Wainer, “without profoundly changing the current item selection procedures *in order for an item pool's security to increase linearly the size of that pool must increase exponentially*” (Wainer, 2000). More concretely, “the strategy of trying to obtain a secure item pool by merely increasing the pool's size is not practical. An eightfold increase in item pool size can be compromised with only a doubling of effort: instead of two weeks it might take three, instead of using 20 burglars they might need 30. In any case, it means that the increased cost of stealing is tiny compared to the increased cost of item pool development. It also means that to maintain the same level of security of a test over time there must be exponentially more items added to the pool. This implies an exponential increase in the cost of test development. It is startling and discouraging to compare the economics of CAT with those of traditional paper and pencil testing. In the latter, linear increases in test volume yield *linear reductions* in the per examinee cost; in the former they yield *exponential increases*” (Wainer, 2000).

Indeed, every aspect of CAT that makes it desirable and preferable to traditional paper-and-pencil (PnP) testing plays right into the hands of “item thieves.” The fact that the item bank test information function (a test information function measures the collective information contained across items comprising the test) reflects the population's distribution of ability means that are fewer difficult items to begin with, and maximization of item information results in

subjects with higher levels of ability being actively corralled towards these items. In short, the aspects of a computerized adaptive test that are beneficial from a psychometric perspective are the same aspects that are easily exploited by agents seeking to compromise the security of the test.

Statement of Problem

The genesis for the to-be-proposed strategy of item selection was substantially motivated by the intention to resolve a long-standing practical issue in testing: item thievery. Wainer (2000) discusses how item theft is remarkably cost-effective as a means to either develop quality preparatory material, or compromise the security of an examination entirely, depending on perspective. The problem is discussed in-depth in the next chapter: literature review.

The strategy for item exposure control proposed here attempts to solve this issue by striving for two general objectives: (1) subvert conventional strategies for item theft and comprising test security, *and* (2) do so making minimal sacrifices in the efficiency of ability estimation (and, indeed, improve ability estimation relative to other modified-CAT methods if possible).

In order to demonstrate how incorporating precision of the ability estimate into item selection can achieve (1) and (2) above, I present the following scenario: there exists some hypothetical item pilferer who can answer questions correctly at will and has perfect memory. In the case of a pure CAT (that is to say, a CAT that is performed with the sole intent of maximizing information in item selection, without regard for other concerns), this individual would quickly steer themselves toward whatever types of items he/she wanted to pilfer, and the test naturally would continually administer such items until either a length constraint or other test

termination condition was met. From the perspective of the pilferer, this is an efficient and rather easy exercise.

This method aims to greatly decrease the efficiency of such a practice so as to render it essentially nonviable. This will be accomplished by using information the subject does not have access to, namely: relative standard error of the ability estimate, item usage frequencies, and projected item usage frequencies based on responses from other subjects.

Ultimately, the goal is to develop a method that does a better job at ability estimation in uniform exposure scenarios by exploiting differential precision of ability amongst subjects.

Chapter II: Literature Review

Initial Development of Item Exposure Control Methods

This section will be a discussion of the myriad different strategies for controlling item exposure. The order of presentation corresponds roughly to the order of development, where older strategies are discussed earlier. This way of presentation is useful in that it allows the reader to glean some of the broader trends in the literature, as well as an understanding of why those trends might have occurred, without in any way diminishing the quality or depth of the information provided.

The early problems with CAT prompted many researchers to attempt solutions, a large number of which were evaluated and categorized in “A Review of Item Exposure Control Strategies for Computerized Adaptive Testing Developed from 1983 to 2005” (Georgiadou, Evangelos & Economides, 2007).

Of the initial proposed mechanisms to control item exposure, many fall within the “Randomization” approach, which from a mechanical perspective simply encourages the selection of slightly sub-optimal items by randomly selecting amongst a large number of candidate items that could reasonably be selected on a CAT. More concretely, one could consider a case where a program assesses the top 5 possible “next items” (where items are ranked based on an information function), and selects randomly amongst them. This process of rebinning available items and randomly selecting from an “approved” set of suboptimal items is called “Randomesque” (Kingsbury & Zara, 1989) and is a prototypical example of the “Randomization” approach.

The actual effect of such a method is essentially distributing what we may consider to be the “responsibility” of the maximally informative item amongst a select number of accepted

alternatives. Naturally, this diffusal of “responsibility” results in a more uniform distribution of item exposure (as seen in Kingsbury & Zara, 1989; McBride & Martin, 1983; Revuelta & Ponsoda, 1998; and others), and is a strategy that scales well with increasing item bank size. A general limitation with the strategy is that purposefully selecting slightly sub-optimal items will affect the precision of the ability estimate, though this can be mitigated to a certain degree with strategies proposed by Wainer (2000), where testing sites can modify aspects of the item bank as dictated by the expected examinees. The example provided was to provide larger numbers of low/normal difficulty items to centers that administer exams for inner-city schools, while providing additional higher difficulty items to sites that service more examinees from charter schools. The larger numbers of items catered to the examinee pool would therefore result in pseudo-randomization strategies selecting comparatively fewer impactful sup-optimal items than otherwise expected -- though it is important to note that the solution in this instance is not the result of the method but rather by modification of the item bank to better suit a different population.

The next wave of item-exposure control strategies discussed in Georgiadou, Evangelos & Economides (2007) are “Conditional Selection” and “Stratified Selection” strategies. Of the numerous strategies that fall under “conditional selection”, Van der Linder & Veldkamp’s Shadow Testing (2005) is particularly well-known, and is an extension of earlier similar work (Armstrong & Jones, 1992; Cordova, 1997; Van der Linden & Reese, 1998). Rather than selecting what the next item might be for an examinee taking a CAT, the algorithm actually generates an entire hypothetical test, and selects the next item from that test to administer. More formally, a linear integer program is formulated to maximize information (for however many

items are remaining), subject to a series of conditions pertaining to a specific examination. Van der Linden & Glas (2010) provide the following example:

$$\begin{aligned}
& \text{maximize } \sum_{i \in O_k} I_i(\hat{\theta}_{k-1})x_i \text{ (maximum information)} \\
& \text{subject to} \\
& \sum_i^I x_i = n; \text{ (test length)} \\
& \sum_{i \in S_{k-1}} x_i = k - 1 \text{ (items already administered)} \\
& \dots
\end{aligned} \tag{1}$$

Where i is an item index, k is current item, n is the length of the test, I is the information function, and “ S_{k-1} is the set of items remaining in the pool after $k - 1$ items have been administered.” The text presentation of this model formulation includes a few subsequent conditions, which have been substituted here as an ellipsis in the interest of expedience.

Implementations of shadow tests involving item exposure controls have historically featured model formulations where a maximum exposure rate is set as a condition the maximization is subject to. Not only is the method effective in controlling item exposure (it can also be used in conjunction with other strategies, like the aforementioned “Randomesque” from Kingsbury & Zara, 1989), but also allows for administrators to ensure that any content constraints are properly met as well as limiting/altogether removing the potential violation of local independence assumptions posed by “enemy items” or mutually exclusive items. Weaknesses with such a method are primarily relegated to the realm of computational complexity, where solving a certain linear integer program quickly is necessary because examinees are themselves subject to time constraints and ideally should not have to wait for the next item to be selected.

“Stratified Selection” strategies are primarily, though not fully, comprised of methods that control item exposure through algorithms that focus on the impact of the a parameter (discrimination in IRT) on the item selection process. Chang & Ying (2001) wrote one of the earlier papers on this topic (and continued their work on this topic further in Chang & Ying, 2008), suggesting that proper item bank usage often means saving highly discriminating items for the latter portions of an exam (a result also indicated in Hua & Chang, 2001). Initial overexposure problems in many examinations were due to the fact that a and information provided are positively associated. This work aimed in part to address this issue, but also had the additional benefit of “saving” the most differentiating items for the later stages of an examination (where the estimate of ability has become sufficiently precise that highly differentiating items are particularly useful).

The final categories of item exposure control strategies from Georgiadou, Evangelos & Economides (2007) are “Combined Strategies” and “Multiple Stage Adaptive Test Designs”. The following discussion will focus on recent papers that attempted Multiple-Stage solutions.

Multi-Stage Tests and Item Exposure

As a baseline, we can revisit the two proposed item exposure control solutions from Wainer’s 2000 paper.

The first can be described simply as “make the test non-adaptive, such that all items are equally likely to be administered”. Certain beneficial aspects of a computerized test are retained (such as immediately available results, flexibility with scheduling, ability to administer tests year-round), though this approach relinquishes the efficiency benefits of CAT and therefore the test would need to be substantially longer (~30-40%, as estimated by Wainer and through other simulation studies) to maintain an equivalent level of precision.

This solution would theoretically at least balance the relationship between item bank size and test security, such that linear increases in cost are met with linear increases in security. From the perspective of maintaining the viability of a computer-based test, this is a plausible solution. However, this balance is somewhat of a Pyrrhic victory: all “adaptive” aspects of a computerized adaptive test are sacrificed to ensure test security. In this light, the strategy might be framed more as complete capitulation rather than a mere concession, however understandable it might be (and, technically, it does constitute a solution to the Zipf’s law component of the item exposure problem).

The second proposed solution in Wainer (2000) involves an item selection algorithm that produces tests above a certain arbitrary threshold of “quality” while meeting specific exposure constraints. An evaluation of the method is perhaps best expressed by the author: “We do not mean to imply that this algorithm, when implemented, will solve all CAT security problems, but it will ease those problems caused by the overuse of a small subsegment of the pool” (Wainer, 2000). In essence, these constraints could allow the production of many sub-optimal CATs, as a general compromise between optimal/fully adaptive versus completely non-adaptive. The unifying characteristic of this solution and all others thus far discussed is that they all result in the creation of sub-optimal CATs, where the degree of suboptimality is assessed typically based on exposure rates and expected bias across the ability continuum.

A multi-stage test takes this idea (creating a small number of tests that meet certain practical and psychometric criteria) by forming testlets and routing subjects to different difficulty testlets based upon earlier performance. In the various methods that will be discussed, these testlets can be created either in real-time during test administration, or created prior to test administration. Often creation prior to administration involves a panel of content experts and a

smaller number of tests, whereas computer-based assembly methods are less subject to certain human limitations that allow them to mass-produce viable tests (assuming proper criteria are indeed specified).

A prototypical multistage design can be seen in Breithaupt & Hare (2007), which proposed a multistage adaptive test (perhaps best visualized as a 1x2x2 panel design, seen in Figure 1) where subjects are routed to harder or normal difficulty item banks based upon performance on earlier questions. Breithaupt and Hare explain that “solutions have been proposed for test construction using either LP (linear program) or non-LP methods”, meaning that there are essentially two classes of assembly methods: methods that are preoccupied with a truly “optimal” solution (given the specification of the LP), and methods that seek at least an acceptable solution. Methods that are performed in real-time during examination tend to be of the latter class, given computational complexity and time available to solve the problem. Shadow tests are a notable example that implement LPs in real-time, and this may be partially explainable now as a difference in computational power. Advancements in modern LP solvers have rendered previously difficult and time-consuming problems trivial to a degree, which addresses this inconsistency (that Breithaupt & Hare implied that LP and real-time are mutually exclusive, when it no longer is).

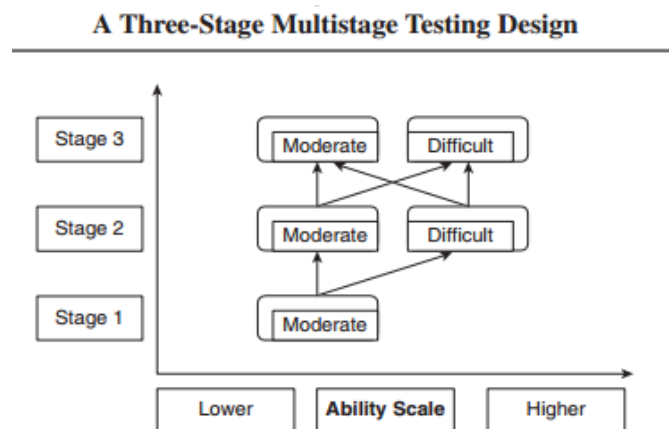


Figure 1. A three-state multistage testing design

The LP vs non-LP explanation is necessary in order to understand one key executive decision in the implementation of Breithaupt & Hare's 2007 method: they decide to use testlets assembled prior to administration (favoring the "optimal" approach). Their method features creation of "hard" and "normal" items, which appear in a ratio of 2:3 (the upper difficulties in the 1x2x2 are "hard", and the lower three item banks are "normal").

Providing a self-evaluation of the outcome: "our empirical analysis using a representative sample of 1,000 candidate ability scores indicated that expected probabilities were always greater for the M testlets. We summed the probabilities for all testlets across the panels and expressed these as the proportion of candidates who would see that testlet or item (not as a probability). When summed, our calculations yielded maximum exposure value of 12% of the candidates scheduled during an 8-week administration, with an average of about 7%. Because items are not shared across the testlets, this exposure estimate for testlets represents the expected exposure of items. These estimates are well below the conventional maximum for high-stakes examinations" (Breithaupt & Hare, 2007).

Without delving too far into the details of the method proposed in their paper (which, though interesting, constitute a digression from the main point here), the authors acknowledge that certain items are still disproportionately exposed (though not in a problematic way according to rules of thumb the researchers cited: 20%, 30%). The smaller number of "hard" items here, relative to "medium" (ratio of 3:2) means that recovery of hard items would probably be relatively straightforward. Certain strategies proposed by Wainer (e.g. increased depth of certain difficulty set for different sites) could be somewhat cost-efficient as a prophylactic measure against "pilfering", but all of this still fails to address the underlying problem of item overexposure (with respect to efficiency of costs for countermeasures, we are still in an

exponential increase vs linear increase competition here). More succinctly, although exposure rates are better-controlled here, the disproportionate item selection frequencies would still ultimately render the method exploitable to item thieves. Additionally, the dissection of results along this single dimension, global exposure rates as compared to conventional maximum exposure rates, suggests that the exposure problem is conceptualized in somewhat simplistic terms. Item characteristics drive these types of solutions, ignoring the wealth of information available using item characteristics in conjunction with examinee characteristics (as an example, examining whether certain subjects of different ability are equally likely to see a different type of advanced *content*, not simply item difficulty).

Along with Edwards, Flora and Thissen (2012), these two papers try to improve upon Lord's "Applications of item response theory to practical testing problems" (1980), which was a further development of his 1971 article "A Theoretical Study of Two-Stage Testing". In this method, a test is separated into an earlier, non-adaptive portion, where subjects are routed to an appropriate subsequent item bank based upon performance on the initial item set. The method can be presented as a viable in-between on the paper-and-pencil versus full CAT continuum; Choi et. al (2010) concluded that although CAT has clear advantages in efficiency of ability estimation when compared to equivalent-length "short forms" of a non-adaptive test, two-stage testing performed "so close to CAT that it warrants further consideration and study".

Criticism of Lord's two-stage approach is levied by Edwards, Flora and Thissen (2012), where the initial concession of "under some circumstances a well-designed two-stage test could achieve nearly the measurement precision of an item selection CAT" is quickly undermined with the suggestion that "the two-stage design may have an Achilles' heel: Due to inescapable errors of measurement on the routing block, some examinees may be administered a second block that

is less than optimal for them. With no third stage to ‘correct’ the single mis-routing, score precision may suffer for those examinees.”

To address this weakness, Edwards, Flora and Thissen (2012) also suggest a multistage adaptive approach, with what is essentially a 3Mx3x3 design (where the 3M indicates three normal difficulty stage 1 forms). Uniform item exposure multi-structure (uMFS) is the name of the method, which also features testlets assembled prior to test administration. The main methodological contribution here is the aim to rigorously control item exposure through the use of routing rules. Figure 2 (below) provides a visual depiction of this test administration method.

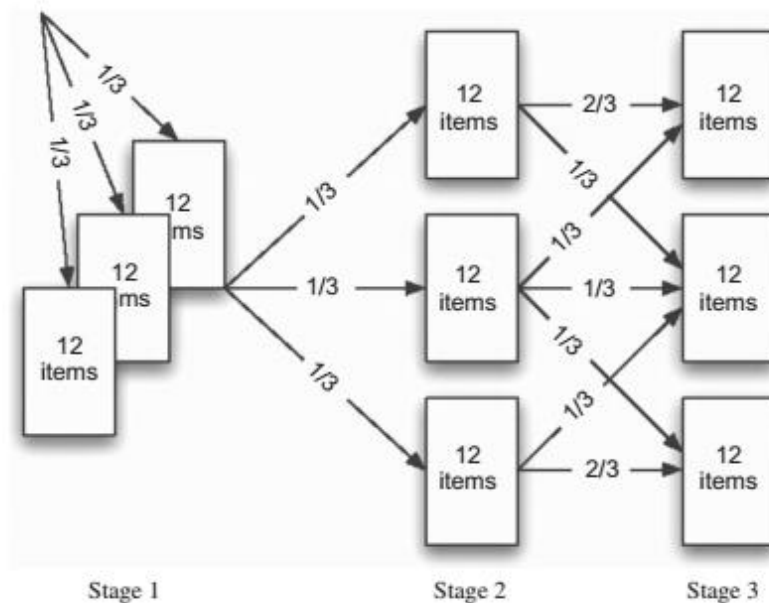


Figure 2. Item exposure control test administration method

On the continuum of paper-and-pencil to pure CAT, this comes much closer to the paper-and-pencil. With pre-defined testlets that are vetted by experts and absolute control of item exposure, the only remotely adaptive components are the two routing nodes. A limitation of this routing system (and, more generally, for any adaptive system) is that the first node is more impactful than the second, in the sense that there are three possibilities for the first and only two

for the second, not to mention the simple fact that the first routing decision determines the terms the second routing decision is made upon. The temporal component of impactfulness in an adaptive examination is of paramount importance. In this examination, an examinee incorrectly routed to either the “easy” or “hard” sections of the exam cannot ever get a set of items that are “hard” or “easy”, respectively.

To reiterate, the authors suggest that “Achilles’ heel” of Lord’s two-stage test is that measurement error can be introduced in the single routing node, such that the test information function (TIF) provided may not resemble an optimal test information function, and argue that adding additional sources of routing is effectively increases the probability that the TIF matches the optimal TIF. Asymptotically this property of routing is obvious, but the practical difference between two and three stages (i.e. routing once or twice) represents what the authors presumably intended to demonstrate is an improvement on this problem with Lord’s two stage, while still having a test of appropriate length.

In their own words, “many exposure control systems reduce or place probabilistic limits on the number of times an item can be used, but most do not require that an item ever be used at all. Absolute exposure control (as we are using that term here) places a firm value on the number of times an item is used, not just a flexible upper boundary” (Edwards, Flora, & Thissen, 2012). The authors additionally discuss two possible implementations, which essentially boil down to (1) incorporating calculations like target information functions (TIFs) and practical cut scores and (2) using a test assembly method from Belov and Armstrong (2008).

Rather than simply place upper bounds on the number of times an item can be used, the concept of absolute exposure control refers to the idea that the number of times an item is selected can be predetermined. This is also comparatively easy to implement when assembling

is performed pre-exam, rather than in real-time. Implementation of absolute item control involves a-priori manipulation of branching rules between testlets (specifically targeting branching cut-scores and target information functions, or imposing constraints during the test assembly process).

uMFS achieves its aims in part using the four equations provided below.

$$avg[se[\theta|x_s]^2] = \sum_{all\ x_s} [se[\theta|x_s]^2] P_{x_s}. \quad (2)$$

$$var[se[\theta|x_s]] = \sum_{x_s} [se[\theta|x_s] - avg[se[\theta|x_s]]]^2 P_{x_s}. \quad (3)$$

Equations (2) and (3) concern the estimate of the latent ability, θ , certain goals related to the precision of its estimate. Equation (2) is a formula for the average error variance in a θ estimate, which is minimized to achieve the most precise estimates. Equation (3) is a formula for the variability in the precision of the θ estimate for all values of θ . This equation is used to ensure uniform precision in the estimation of θ across its entire continuum.

$$var[l(\theta)] = [QI]^{-1} \sum_{q=1}^Q \sum_{i=1}^I [I_i(\theta_q) - \bar{I}(\theta_q)]^2. \quad (4)$$

Equation (4) relates is the testlet information function for the first stage of the exam administration (which in this study is administered non-adaptively). Computing variance of the the testlet information function allows one to determine the degree to which the testlets provide equivalent information. This function is therefore minimized to obtain testlets with equal or nearly equal information functions.

Finally, equation (5) relates to the proportion of examinees routed to each subsequent testlet, with a value of zero representing equal proportions (and, consequently, optimal conditions for controlling item exposure).

$$var[prop_{ij}] = [ls]^{-1} \sum_{j=1}^s \sum_{i=1}^l \left[prop_{ij} - \frac{1}{l} \right]^2. \quad (5)$$

Edwards, Flora and Thissen (2012) subsequently operationalize these criteria in the form of the weighted linear combination

$$crit = wt_1 avg [se[\theta|x_s]^2] + wt_2 var [se[\theta|x_s]] + wt_3 var[I(\theta)] + wt_4 var [prop_{ij}]. \quad (6)$$

where wt_1, \dots, wt_4 denote weights placed upon each criterion. Differential weighting of the criteria can be used to produce exams that provide optimal item exposure control, optimal accuracy and precision in ability estimation, and everything in between.

As it was the Breithaupt and Hare (2007) paper, routing is performed based on number of items correctly answered, rather than the current ability estimate. With the acknowledgement that the estimate of ability will likely be imprecise after only a few items have been administered non-adaptively, it is nevertheless unlikely that any estimate of latent ability at that juncture would be worse than a summed-score. The summed-scores are surprising especially in the scenario where the authors suggest that (due to the method's not truly 3Mx3x3 nature) that the first testlet may justifiably be made longer. Estimated ability methods for latent constructs (EAP, MAP, etc.) at the very least take advantage of the carefully-calibrated psychometric characteristics of the testlets. The computational burden of implementing such a suggestion is so trivial (especially given that this paper was published in 2012, where an EAP calculation on a modern computer takes a fraction of a second) that one may argue that the use of summed scores is an unnecessary limitation, though naturally there are circumstances where EAPs are not ideal

or that summed scores are likely to produce essentially the same results (e.g. as it was implemented here, with parallel medium-difficulty tests to start, where the summed scores of items with highly similar difficulty parameters should closely match the results from a latent score calculation). This limitation is pertinent in the sense that summed scores technically introduce an additional source of imprecision outside of the linear combination (equation 5) used to conduct the routing process.

The authors also discuss how the assembly algorithm saved items with higher a parameters for later on without explicitly being told to do so. This may be reasonably interpreted as the assembly algorithm supporting Hau & Chang's (2001) work, which stated that reserving highly discriminating items for later in the test can be a good strategy for high-quality ability estimation with better than some other methods' exposure control. An interesting finding that will be discussed further in the next segment, 'item influence'.

Item Influence

Influence, in this context, refers to the degree to which an item or series of items has an effect upon the estimation of θ as a result of the determination of which subsequent items are chosen. These have been referred to in this paper already as routing decisions, and are of interest insofar as they have a disproportionate effect upon ability estimation, such that performance on earlier items is comparatively more impactful on the final estimate of θ . Therefore, much of the research in this field relates to which items and what types of items should be presented earlier to later in a CAT.

In a continuation of the discussion of uMFS (above), one of the key findings related to influential items in CAT is that, perhaps counterintuitively, items with higher discrimination parameters should appear later, rather than earlier, in an adaptive test. This recommendation has

appeared both in item-by-item and testlet versions of CAT. Edwards, Flora and Thissen (2012) state “it is interesting to note that the algorithm we employed saved the highly discriminating items until the final blocks... Hau and Chang (2001) addressed the issue of where to place the more discriminating items for maximum benefit in a CAT while conducting research on their method of exposure control (the ascending a-stratified method). The conventional wisdom up until that time had suggested using the most discriminating items early in a CAT. Hau and Chang’s work suggested that some of the gains realized in CAT come from using the most discriminating items at the end of the test. The algorithm we employed placed the more discriminating items at the end of the test (see Figures 3 and 5) even though it was never explicitly instructed to do so. These results give further support to the conclusions of Hau and Chang.”

Hau & Chang (2001), as well as Zhou & Reckase (2014) and others, indicate that inclusion of items with lower discrimination parameters can improve the accuracy and precision of ability estimates in CAT. Zhou & Reckase (2014) specifically discuss this issue in stating that “while highly discriminating items are usually desired, including items with varying discriminating parameters can achieve the measurement accuracy as expected when the distribution of item difficulty matches the ability distribution of the target examinee population”.

Wainer, Chang, Ying, and many other researchers have noted that earlier iterations of the CAT GRE selected highly discriminating items for the initial items in a way that damaged the validity of the test (referred to as the “2000 GRE Incident” in Chang & Ying, 2008, also mentioned in Zheng & Chang, 2015). More specifically, if an examinee were to perform poorly on these disproportionately weighted early items, there were no measures (either via a variable-length test or simply a longer test) that the participant’s true ability score could be recovered.

The mechanism by which such a phenomenon occurred was that participants with poor early performance were subsequently presented with a barrage of low difficulty and low discrimination items, creating an abyss where even excellent performance could help the participant crawl out to achieve a reasonable estimate of his or her ability.

A natural correction for this type of problem is the implementation of a variable-length test: an idea that has been proposed in a few other contexts, notably with mastery examinations. For a mastery examination (e.g. whether someone is sufficiently qualified to enter a medical profession, or to practice law), some essentially arbitrary threshold is drawn that divides the latent ability continuum into a region for “pass” and a region for “fail”.

A logical consequence of using an adaptive test with a certain threshold set on the ability continuum is that participants who quickly demonstrate that they are not near the threshold (i.e. clearly fall within the “pass” or “fail” regions) will have shorter tests. A sufficient level of precision to estimate ability is reached, the test makes a determination that the examinee falls into one of the two categories, and the test ends.

As participants get closer and closer to the threshold that divides “pass” from “fail”, increasingly precise estimates of ability are necessary to determine (within a reasonable degree of confidence/certainty) which category the subject belongs to. One can imagine a particularly extreme case where an individual is so very, very close to the threshold of passing (but nevertheless is truly, according to ability, a deserved fail) that a nearly endless barrage of items that are highly discriminating with difficulty near the cutoff region are necessary to make a final determination on that subject’s result. Many of these ideas are expounded upon in Vos & Glas (2009)’s discussion of “*sequential mastery testing*.”

From the perspective of fixing earlier iterations of the CAT GRE, however, a variable length test would be useful because participants who fall into the initial hole that previously was impossible to fully escape from could drag themselves out when presented later opportunities. The precision of their ability estimates will likely be lower than those who truly belong at the lower end of the ability continuum: this is a metric by which the next step-- expanding the item selection routine to allow for items deviating farther from the present EAP or MAP-- can be justified. Correct answers on those items would further widen the current distribution of the subject's ability (and shift the mean and/or mode of the posterior ability distribution), such that eventually a participant could eventually be presented with difficult questions: another opportunity to get closer to the subject's true ability. This is of course a rather roundabout way of getting to those difficult questions ("Why not just do well from the start?"), but psychological factors like "computer familiarity, facility, (and) anxiety" (Zwick, 2010) are relatively well-established and therefore, in the interest of estimating ability as accurately as possible, it is necessary, in some form or another, to have a mechanism wherein poor early performance can be remediated through high performance later on.

The initial GRE approach of front-loading highly discriminating items is evaluated via simulation in Chang & Ying (2008), which supports the interesting observation that maximizing blindly based upon Fisher Information provided by a single item can actually produce suboptimal results. "Our analytical derivation shows that, under the maximum information item selection strategy, if an examinee failed a few items at the beginning of the test, easy but more discriminating items are likely to be administered. Such items are ineffective to move the estimate close to the true θ " (Chang & Ying, 2008). This assertion was further supported empirically via simulation, MSE and bias of estimates of ability were near-zero for ascending-*a*

item selection but systematically biased with higher MSE for descending- a item selection (with increasing severity as initial estimate of ability deviates further from 0 on $\sim N[0,1]$).

Additionally, the simulation demonstrated precisely the phenomenon that allegedly fueled the “2000 GRE Incident”, replicating rather clearly the result where true ability could not be reached after poor performance on a few disproportionately weighted early items.

This phenomenon is not unique to the pure CAT version of adaptive tests: Kim and Moses (2014) examined a related concept in the context of multistage testing -- now a term established in the nomenclature as “misrouting.” Their aim was to attempt to understand “the extent to which the examinees' scores would change if they received a module that did not match their ability” (Kim & Moses, 2014). The perhaps surprising results of their simulation revealed that “the impact of misrouting was minimal”, where “the differences between the target and off-level scores were also trivial for practical purposes” and any large score differences or variability occurring at either the low or high region of the scale are not likely to be of practical interest, because differences in those regions will rarely result in inaccurate assignments in actual testing situations.” With the caveat that “examinees’ final scores would change depending on the choice of routing”, they state that “the magnitude of score change caused by any systematic routing error will be comparable with the score changes caused by random measurement error.”

If true, the conclusions derived from Kim & Moses (2014) have rather interesting and far-reaching implications for test assembly and item exposure control strategies. Their results indicate that you can purposefully misroute subjects and have limited effect on ability estimation in certain circumstances. Additionally, purposeful misrouting is less harmful for subjects near average ability, so selectively misrouting subjects becomes viable.

Of recent methods in publications that might be well suited to exploit this dynamic is Zheng & Chang's (2015) "*On-the-Fly Assembled Multistage Adaptive Testing*" (OMST, depicted in Figure 3). As described by the authors, OMST "is administered in stages and only adapts between stages... where (instead of having) the modules in every stage... pre-assembled before administration, the stages in OMST are assembled on the fly." The implications of having the "on-the-fly" component here are that (a) testlets do not have to be the same length and (b) there can be variation in number of stages while still administering the same number of items. What this effectively means is that the latter part of the test can be used to administer items in increasingly small testlets (or by themselves) as a mechanism to exert greater control over item exposure after a certain level of precision in the ability estimate has been achieved. Many of ideas presented earlier in this paper could be used to implement a variation of OMST that allows the best of both worlds: full item exposure control and accurate, precise ability estimation.

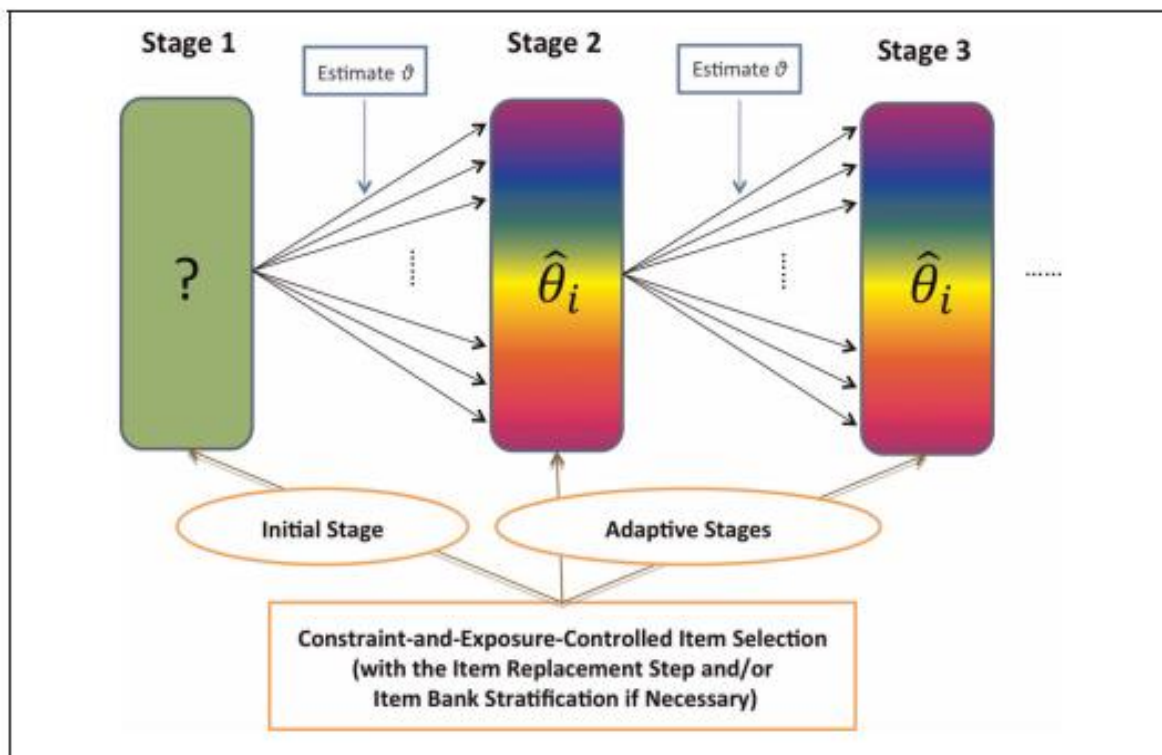


Figure 3. On-the-Fly assembled multistage adaptive testing

Best of all, perhaps, is the fact that the automated assembly removes some additional cost for the test provider, with the acknowledgement that development and maintenance of an increasingly complex, real-time test administration program is no triviality. This kind of complicated test administration program is a natural segue into a discussion about item bank management, where the focus is on the overall characteristics and composition of items within the item bank. In order of OMST to work properly, as an example, one would need to continually develop and vet new items so that over time item parameters do not drift due to overexposure. More plainly, an item that becomes well known is one that has its parameters “drift” because it has become (typically) easier for reasons unrelated to the actual subject matter contained in the item, so proper bank management is important in evaluating that such an item may need to be rewritten or rotated out of the available item pool.

Item Bank Management

Another subtopic within the item exposure literature is the management and maintenance of item banks. In simplest terms, an item bank or “item pool is a collection of test items for a given domain usually stored in computer memory along with a list of codes for their attributes. All item attributes can be classified into three different categories, namely categorical attributes (e.g., item content, cognitive level, and item format), quantitative attributes (e.g., word counts, exposure rates, and item-parameter values), and logical attributes (e.g., relations of exclusion or inclusion in subsets of items) (van der Linden, 2005, chap. 2)” (a self-reference from Adelaide, van der Linden, & Veldkamp, 2006).

Of particular importance in the management of item banks is the idea that the quality of an item bank “depends only partially on the distribution of the item parameters” (Adelaide, van der Linden, & Veldkamp, 2006). These authors, as well as some soon-to-be-mentioned others,

indicate that due to additional variables introduced during test assembly/administration like content constraints and adaptive item selection, the “key to the performance of a test-assembly algorithm is the match between the composition of the pool and the complete list of specifications for the tests assembled from it”. In other words, a good item bank is one from which many tests meeting a certain set of specifications can be derived, such that test administration can be free from security concerns while simultaneously achieving the stated objective of the test.

A more concrete example of how an otherwise well-constructed item bank (from the “quantitative attributes” perspective) can go awry: “in spite of their quality, highly discriminating items are not chosen if their content attributes are not the ones required by the test specifications. In an extreme case, if the pool misses one item with a specific combination of content attributes, it can even become impossible to select a valid test from the pool at all. Also, items in the pool with relatively scarce combinations of content attributes are easily overused. The latter problem creates potential item-security risk in computerized adaptive testing and leads to the necessity of an adaptation of the item-selection algorithm to prevent items from being overexposed” (Adelaide, van der Linden, & Veldkamp, 2006).

These authors touch upon one of the deceptively important aspects of item bank management, which is often not represented in papers evaluating the viability of new adaptive testing methods: circumstances where incongruence between the composition of an item bank and the content requirements for tests using that item bank are often where item exposure problems arise. The “categorical” and “logical” attributes mentioned above represent new salient features in the problem-space of managing item exposure, though are likely outside the bounds of what will be necessary to pursue the intended research problem.

One practical aspect of item bank management that ties many of the ideas discussed so far together is the way that the ratio of the number of items administered divided by the total number of items in the bank plays out. If such a ratio is small, e.g. 2, then after two uses any method that wants uniform exposure can no longer use that item. In this circumstance, you will see the fastest deviations in item selections for CAT versus methods that incorporate exposure controls. For example, if one had 1,000 items in the bank, a test of length 20, and 100 subjects -- 2,000 total items are to be administered, where uniform exposure conditions would require each item to be used only twice. A CAT will exceed that cutoff of 2 almost instantly, and will almost certainly exceed it many times over before an exam is over. Conversely, if the previous example featured 1,000 subjects, then the new cutoff for uniform exposure would be 20, which would take comparatively longer to exceed, and as such we might expect that CAT and exposure control methods would diverge comparatively later. With this in mind, we can discuss this and other ideas that are relevant to the proposed method.

Chapter III: Methodology

Proposed Method

The reader might correctly determine certain statements contained within this section to be speculative in nature. These statements do not form the core of the argument and general logic in the formulation and implementation of the proposed method, but serve to augment the discussion in an informed manner and to highlight potential difficulties arising from the complex, interwoven set of variables that collectively comprise the system that the proposed method operates within.

This project is primarily motivated by what appears to be an unaddressed and as-of-yet unexploited component in the computerized adaptive testing process: differential precision in estimated ability across participants at various junctures prior to test termination. As discussed previously, there are methods that create soft or hard rules on maximum exposure rates, or through some other mechanism attempt to achieve uniform item exposure: the unifying strategy through which uniform exposure is achieved is the conditional selection of sub-optimal items.

The architecture through which this conditional is defined varies greatly: stochastic item selection after selecting a subset of viable items (e.g. randomesque), deterministic item selection based on predefined criteria (e.g. shadow tests with maximum exposure rates), decision nodes in a smaller network of paths relative to a normal CAT (e.g. multistage testing), and even methods that change what constitutes an optimal item in the first place (minimum expected posterior variance). The common element here is that subjects are not compared directly with each-other or with another hypothetical subject that generally represents a typical subject. At any given point in the testing we already understand that examinees can have not only different levels of

estimated ability but also differential certainty in our estimate of that ability. This latter dynamic in the testing process is the novel component in this item exposure control strategy.

The implications of using the precision of the ability estimate relative to other subjects can be far-reaching if so desired. As an example, one may take the simple operationalization of this concept as a slight modification of a shadow test with maximum item exposure constraints.

Equations in figure 4 demonstrate a linear program (a shadow test, in this terminology) modified to meet this method:

$$\begin{aligned}
 & \text{maximize } i_1x_1 + i_2x_2 + \dots + i_kx_k \\
 & \text{subject to} \\
 & x_1 + x_2 + \dots + x_k = m \\
 & x_1 \dots x_k \in \{0, 1\} \\
 & Q_1 \dots j-1 \cap Q_j \dots J = \emptyset \\
 & \hat{f}_1x_1 < E\{f\} \cdot \phi_n \\
 & \dots \\
 & \hat{f}_kx_k < E\{f\} \cdot \phi_n
 \end{aligned}$$

Figure 4. Linear program of maximum item exposure constraints

The x variables denote item selection; if x_i , for example, equals 1, then item one has been selected by the algorithm (the index for x reaches its end at k , which is the total number of items). In the first constraint, m is a positive integer used to determine the number of items selected in candidate solutions for the algorithm. The subsequent two constraints are formulated such that the LP solver does not produce solutions that are nonviable in the practical sense; x , or a single item, must be used either once or not at all, and items administered to an examinee earlier in the examination are not eligible for re-administration (using Q , indexing from the first item to the current item, j , we say that these sets of items are disjoint regardless of whether $j=J$, where $J = (j-1)+m$). The last constraints essentially state that once a maximum exposure

criterion, $E\{f\}$, is reached, the item is no longer eligible for administration (though this is subject to ϕ , which we discuss in the coming paragraphs). In their totality, the constraints here prevent items from being used more than once (either through re-administration or within a candidate solution), prevent fractions of items from being selected (this is an obvious but nevertheless important part of the LP formulation), and satisfy a pre-specified maximum exposure criterion (modified dynamically through ϕ).

The maximization component selects items to maximize summed information, with constraints ensuring that the shadow test creates a test of length m , that each item is only used once or not at all, that no items are re-administered from earlier in the test (disjoint sets), and that no item usage frequency exceeds the new criterion. This is documented more thoroughly in figure 5 (below):

$$\begin{aligned}
i_{jn} &= I_j(\hat{\theta}_n) = a_j^2 \cdot \frac{(P_j(\hat{\theta}_n) - c_j)^2}{(1 - c_j)^2} \cdot \frac{1 - P_j(\hat{\theta}_n)}{P_j(\hat{\theta}_n)} \\
P_j(\hat{\theta}_n) &= c_j + \frac{1 - c_j}{(1 + e^{-a_j(\hat{\theta}_n - b_j)})} \\
\hat{f}_j &= f_{j_{obs}} + f_{j_{exp}} \\
E\{f\} &= (N \cdot J)/k \\
\phi_n &= SE(\hat{\theta}_{jn})/E\{SE(\hat{\theta}_j)\} \\
\hat{\theta}_n &= E\{\hat{\theta}|X\} = \frac{\int_{-\infty}^{\infty} \theta L(\theta|X) g(\theta) d\theta}{\int_{-\infty}^{\infty} L(\theta|X) g(\theta) d\theta} \\
L(\theta|X) &= \prod_{i=1}^{j-1} P_i(\theta)^{x_i} (1 - P_i(\theta)^{1-x_i}) ; x_i \in \{0, 1\} \\
g(\theta) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}} \\
SE(\hat{\theta}_n) &= \sqrt{\sum_{i=1}^{j-1} I_i(\hat{\theta}_n)}
\end{aligned}$$

Figure 5. Additional functions necessary for the calculations in Figure 4.

Here we have the information function and probability function for a 3PL IRT model, expected frequency of item use being a sum of frequencies of already used items with items that may be used (in some implementations, discussed later). The figure ends with commonplace functions for estimating latent ability, calculating the likelihood of a response pattern using marginal maximum likelihood, the density function for a standard normal distribution, and the formula for the standard error of the ability estimate (as implemented in this study). The objective of this figure is to document supplementary functions necessary in the calculations in the preceding figure. The letters i and j are used to index items here, where i indexes as part of a combined operation with an ordered sum or multiplication, whereas j is used to designate the juncture, represented by item number, to which an examinee has progressed in the examination.

This criterion using ϕ is the primary difference between this formulation and the standard “shadow test” formulation; the maximum exposure rate in the solution for any individual shadow test is multiplied by the ratio of the standard error of the ability estimate divided by the expected standard error of the ability estimate. Practically this means that if one’s standard error of ability is double what might otherwise be expected at this stage of the test, we relax constraints on maximum exposure so as to provide a more informative item (thus attempting to coerce a sort of equality in the standard errors by providing differentially informative items). Conversely, if one had a standard error half that of what is expected, then maximum exposure rates are proportionally constricted. This almost necessarily results in the administration of a decidedly suboptimal item, which we posit is not damaging in this context because we are comparatively more confident of this subject’s ability estimate than estimates of ability for other subjects. To put it simply, when one needs an informative item the method accommodates this need. When an informative item is not needed, the dictates of ensuring uniform item exposure are followed.

On a side note, a worthwhile practical detail that arises from this implementation is that the absolute maximum exposure rate for any given item can be controlled by constraining the bounds of ϕ .

The far-reaching possible extensions are if we not only account for current exposure rates, but also projected exposure rates as dictated by each subject's shadow test. As such we differentiate from observed and expected item exposure (demonstrated earlier with f_{obs} and f_{exp}) and can take each into account in item selection. A practical issue in the implementation could be that items expected to be given are treated as equally influential in the item selection process as items already given. This means that the shadow test for each subject (if no modification is made or this issue goes unnoticed) reflects simply the present precision of the ability estimate and not the projected precision of the ability estimate (which is indeed estimable using the shadow test and the current ability estimate). The effect of failing to account for this phenomenon would be to render expected item exposure rates somewhat useless, as they are dictated by shadow tests that no longer serve as reasonable projections of what items might be administered to a subject. More concretely this means that some subjects would be administered unnecessarily suboptimal items because other subjects in a similar ability range have items the original subject needs contained within their shadow tests -- reserved unjustly, so to speak. The simplest way to solve this issue is to simply not use f_{exp} in the calculation.

There is also the potential issue of computational complexity, where we must recognize certain pragmatic concerns (producing a solution near-instantly for real-time testing scenarios, as discussed by van der Linden, Ariel & Veldkamp, 2006). Though this will be subject to some scrutiny in the to-be-proposed pilot simulation, one might reasonably expect that calculating new shadow tests for all subjects every single time someone answers a question is outside the bounds

of what current technology can quickly perform. I will discuss this issue further in my presentation of the different versions of the proposed method in the pilot simulation, but the general statement to be made here is that this strategy (by virtue of attempting to take into account much more information in the item selection process) can likely easily have its manifestations rendered nonviable if specified without consideration for current computer hardware and LP solving software.

Proposed Method: Caveats and Justifications

The introduction of a novelty often initially requires an arbitrariness of sorts, in the sense that the operationalization of some broader concept has to start somewhere. In this case, we encounter the difficulty of operationalizing mathematically the "target reliability" and "relative reliability" required to calculate the need a simulated examinee has for an informative item, though earlier discussion of differential levels of precision in the ability estimate among subjects appears to be a natural place to base an initial definition. We take the relative reliability of an ability estimate as a ratio, where the standard error of a subject's ability estimate is compared to the expected standard error of the subject's ability estimate (i.e. "target reliability"). This method can later take the simple form of a multiplicative change in the maximum allowed exposure rate in the formulation of a LP, as seen in figures 4 and 5.

How does one define this expectation, though, especially when in this context it refers to a group-level characteristic? A simple way to do this is to take the mean standard error of ability for all subjects at any given stage in the test -- this can be performed either in real-time or a set of expected standard errors can be obtained through earlier simulations (this is a simple consideration for a simulation, but the practical implications regarding implementation render this an important decision that will be expounded upon later). This definition has several

advantages: (1) the calculation is simple, (2) inherently strives to homogeneity of variances for subjects in all areas of the latent continuum, and (3) is easily implemented in the form of a multiplicative change in maximum allowed exposure rate (with constraints on the multiplicative changes if so desired, as mentioned earlier). It is worth noting that if performed in real-time, one has the additional difficulty of differential speededness in test progression. In a simulated environment this can be rendered wholly irrelevant, but in application one might reasonably decide that obtaining expected precisions pre-test via simulation is optimal.

In addition, should the expected precision be based on the method of implementation or a hypothetical CAT with no item exposure constraints? If it was based on a pure CAT, one would continually be relaxing maximum exposure constraints because a disproportionate number of subjects would have higher need for an informative item. Given that a pure CAT maximizes information, the only way to consistently meet the expected precision for all subjects would be to administer items as if the method were a pure CAT, rendering item exposure considerations subordinate to the necessity of keeping pace with respect to precision. As such, expected precision probably should be dictated by averages within the method itself, rather than benchmarks calculated in other methods.

Software

Data simulation and analysis will be performed using R. Specifically, packages “catR” and “gurobi” will be used to solve CAT/IRT and linear programs, respectively. Post-analysis data-processing will also be performed in R.

Pilot Study

A pilot study will be performed in order to determine which permutation of the proposed method best meets the needs for the primary simulation and application (effectiveness and speed

are the main considerations). This pilot study will feature a panoply of other item exposure control methods to determine which are functionally redundant, with the intention of representing different types of strategies (for the sake of later comparison) efficiently. The proposed methods will be labeled Nonadaptive (fixed-form: it is intended to mean a totally nonadaptive test, as Wainer 2000 suggested as a solution to this problem), Randomesque-5, Rebinning (allow best item within a subset of acceptable items that have been classified as underexposed), Shadow Test with maximum exposure constraints, Need-Adjusted Shadow Test without expected frequencies, Need-Adjusted Shadow Test with expected frequencies, and of course a “normal” CAT. These are intended to adequately represent the full continuum of Nonadaptive to CAT.

Item parameters in the pilot simulation can be seen in the table below:

Table 1. Item parameters in the pilot simulation

<i>Parameter</i>	Distribution	Distribution Prop. 1	Distribution Prop. 2
<i>A</i>	Uniform	.5	2
<i>B</i>	Normal	0	1
<i>C</i>	Uniform	.1	.25

Other parameters will be varied to reveal differences between the methods (table 2, below).

Table 2. Other parameters included in the simulation

Sample Size	100, 1000, 10000
Test Length	30, 40, 50
Item Bank Size	100, 1000, 10000

After a proper selection of methods is determined, the main study will commence.

Primary Study

The primary study will compare the remaining methods with the selected implementation of the proposed methods across a variety of outcomes and a simulation condition varying the size of the item pool. Table 3 provides information on simulation conditions, and Table 4 provides information on to-be-evaluated outcome measures. Additionally, certain parameters in the item generation process were modified relative to the pilot study, as seen in Table 5.

Item parameter distribution selection was performed with the intention of generalizing to a somewhat “generic” four-response multiple choice item bank, while still maintaining appropriately varied possibilities in draws from the distributions that the differences in the tested methods are coaxed out. Notable is the added decision to generate the parameters from independent distributions, such that a , b , and c are unrelated. This may differ in practice from certain specialized types of item development where these parameters might otherwise be reasonably expected to be related.

The a parameter, distributed uniformly on $[0.5, 3.0]$, is distributionally similar to many substantive applications of IRT where, out of necessity and the directive to efficiently recoup item development costs by using as many items as possible, a broad range of potential values for the discrimination parameter are possible. A more concrete detail arising out of this distributional decision is that, as a consequence of the item information function in the 3PL model, the 0.5 to 3.0 range represents a corresponding 0.25 to 9.0 multiplicative change in the information provided by an item. As such, from the standpoint purely of item information, significant variability in the resultant value will be induced through the a parameter.

The b parameter is normally distributed, specifically $\sim N(0, 1.25)$. The expanded dispersion is in accordance with item bank designs that intentionally induce a slight

incongruence between the population ability distribution, theoretically distributed as standard normal in many cases, in order to generate more broadly useful anchor sets or initial parallel forms (often discussed in terms of “mini” or “midi” tests).

The c parameter is distributed uniformly on $[0,0.25]$, representing four-response multiple choice items with a single correct answer and varying quality of distractor responses. The c parameter, much like the a parameter, will generate substantial variability in values from the 3PL item information function for items generated in this simulation.

Table 3. Information on simulation conditions

<i>Methods Used</i>	CAT, both types of Shadow Test, Fixed-Form
<i>Sample Size</i>	750
<i>Item Bank Size</i>	500, 1000, 1500
<i>Test Length, Number of Replications</i>	40, 35

Table 4. Information on to-be-evaluated outcome measures

<i>Bias of Ability Estimate</i>	Empirical Densities, Tables if Bias exists
<i>Properties of Standard Error of Ability</i>	Mean, Range, Distributional Properties, Relationships with other variables or methods
<i>Item Exposure Control</i>	In aggregate, as well as by parameter in 3PL IRT model
<i>Item Usage Overlap</i>	Methods vs. CAT, comparisons involving high-ability subjects

Table 5. Modified parameters in the item generation process

<i>Parameter</i>	Distribution	Distribution Prop. 1	Distribution Prop. 2
<i>A</i>	Uniform	.5	3
<i>B</i>	Normal	0	1.25
<i>C</i>	Uniform	.0	.25

Pilot Study Results

The pilot study was used to make informative decisions regarding parameters and methods to incorporate into the main study. Selected representative figures will be presented to support the conclusions obtained through this pilot simulation

Ability estimation was unbiased in all methods and, as seen in figures 6 and 7, pure CAT reigned supreme with precision-based shadow tests and uniform exposure shadow tests falling slightly behind. Fixed-Form (transiently referred to as PnP, or paper-and-pencil) was easily the method with least precise ability estimates, and therefore a potentially useful foil to pure CAT.

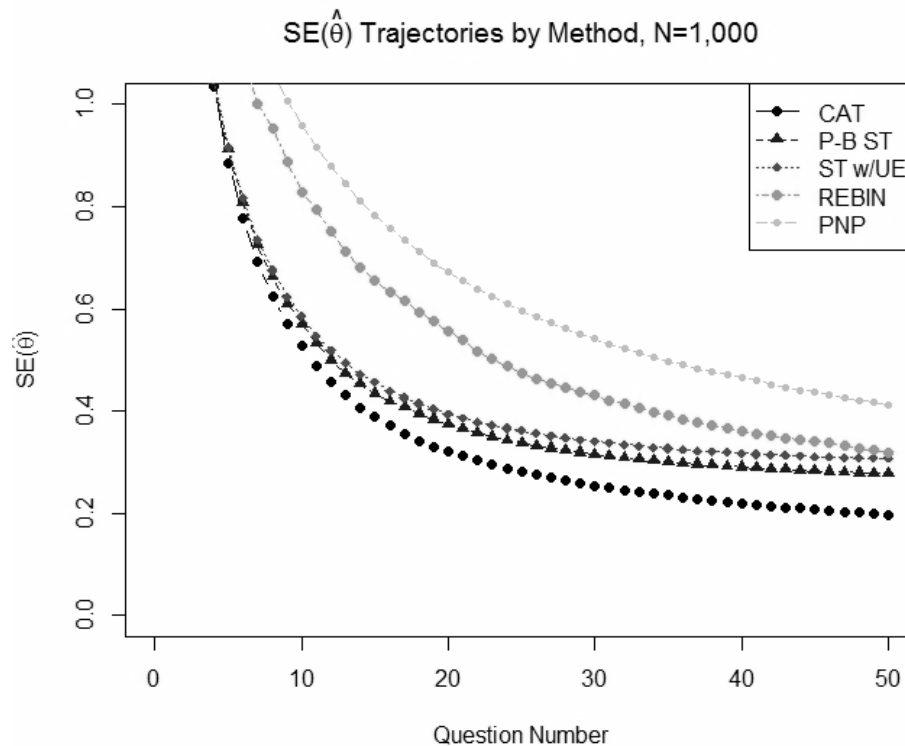


Figure 6. $SE(\hat{\theta})$ trajectory by method in pilot study results ($k=1,000$)

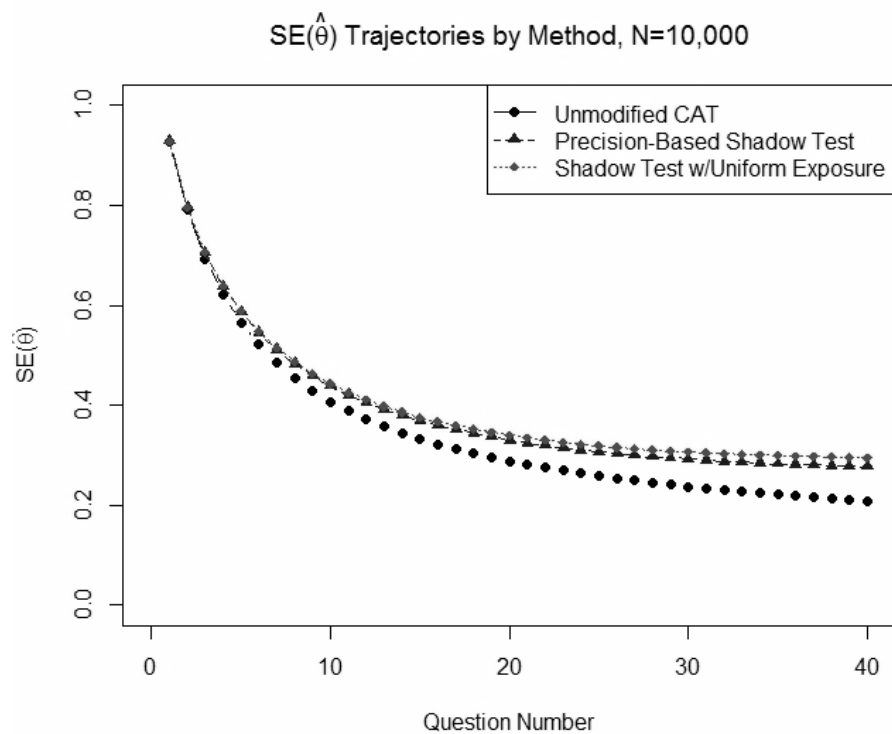


Figure 7. $SE(\hat{\theta})$ trajectory by method in pilot study results ($k=10,000$)

Item exposure control, as seen in figures 8 and 9, also largely featured results that one might expect. Uniform exposure shadow tests performed unsurprisingly well. Precision-based shadow tests appear to perform well contingent on multiple factors, including depth of item bank, and as such circumstances with smaller item banks appear to create particularly uneven exposure rates.

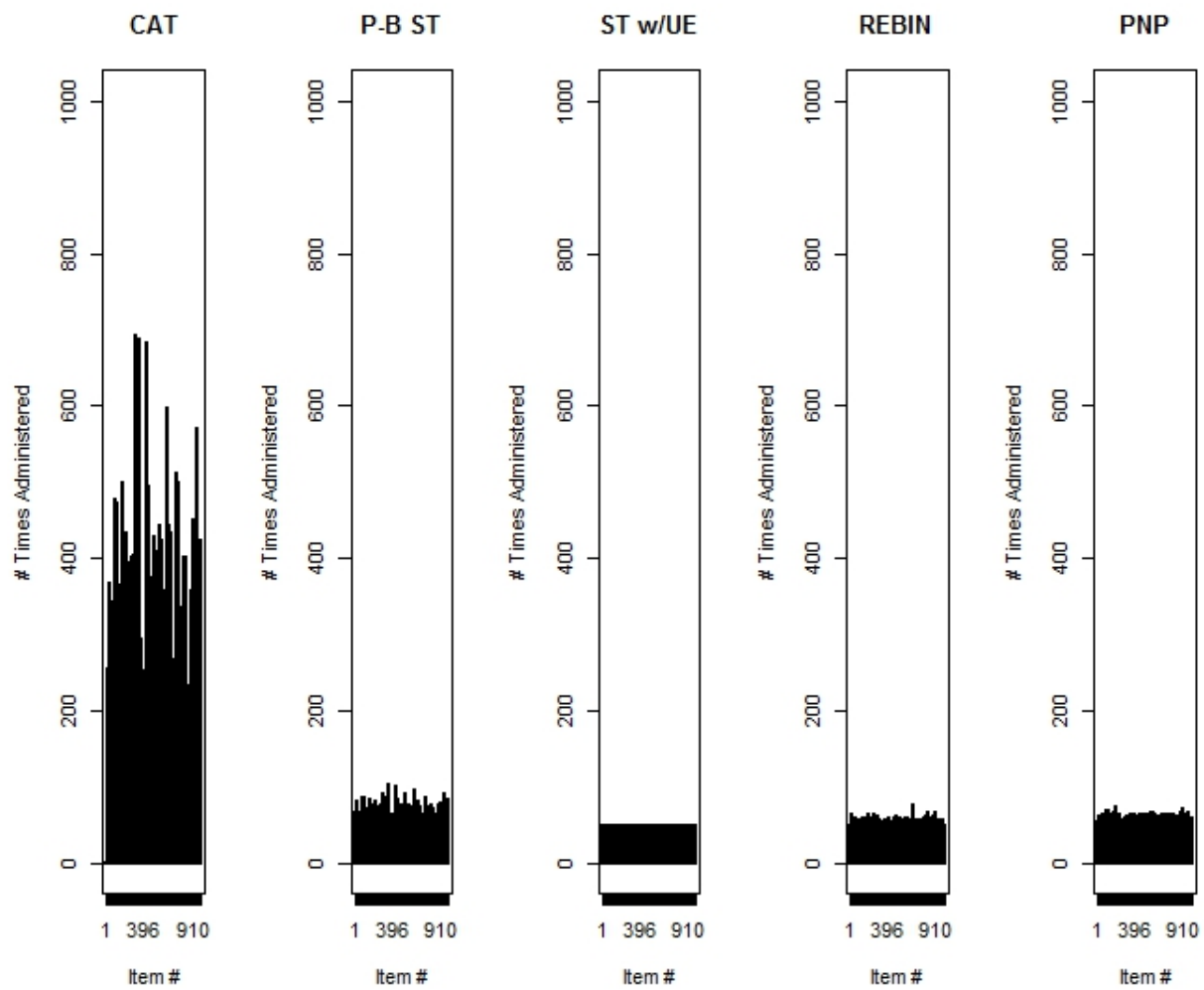


Figure 8. Comparison on item exposure control by method

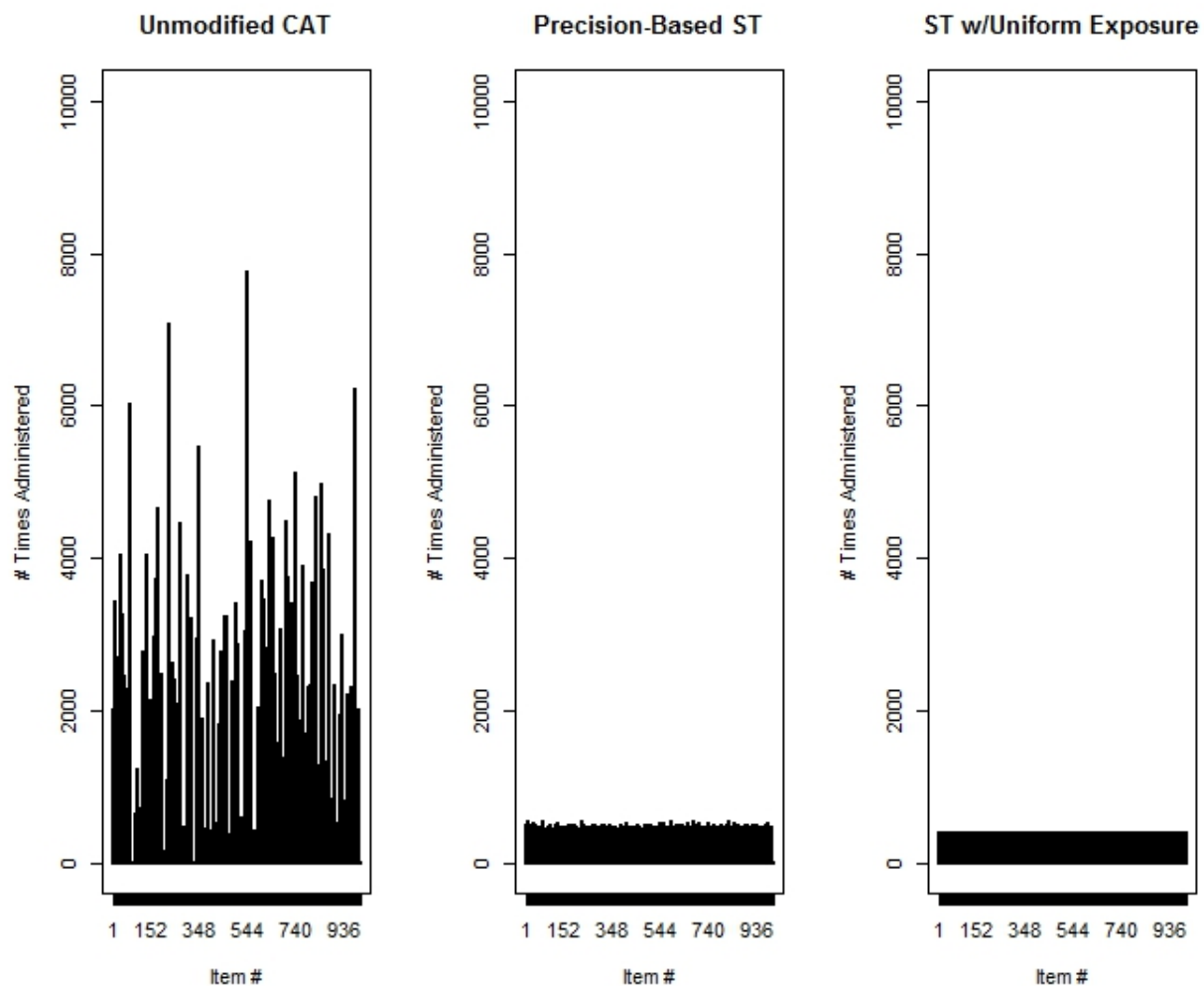


Figure 9. Comparison on item exposure control by method

Item usage overlap was calculated for precision-based as well as uniform exposure shadow tests as an initial way of conceptualizing the degree to which the methods might avoid item pilfering problems that historically plagued pure CAT. As seen in figure 10, the limited overlap for parallel subjects in item selection by method is a promising result that will require further investigation in the primary study.

The major conclusions derived here are as follows:

1. CAT will be used as an exemplar for ability estimation.
2. Nonadaptive item selection will be used, fulfilling the need for a method at the entirely opposite end of the adaptive testing spectrum (a non-adaptive test that will weakly converge in an asymptotic manner on uniform item exposure, rather than deterministically).
3. “Rebinning” is sufficiently similar to uniform exposure shadow tests that it is redundant and will not be included in the main study.
4. Information pertaining to computation time and other pragmatic concerns was used to inform conditions in the primary simulation. Principally, the implementation of the precision-based shadow test must account for certain real-world environmental variables inherent to high-stakes testing formats (i.e. due to a multitude of factors, computation time for item selection must take at most one second, though ideally less). As a result, the shadow test implemented here does not have a “shadow” in the standard conceptualization of the method: it will be of length one. The method retains its name, but is fundamentally a form of conditional subsetting contingent on how the integer program in the item selection algorithm is formulated.

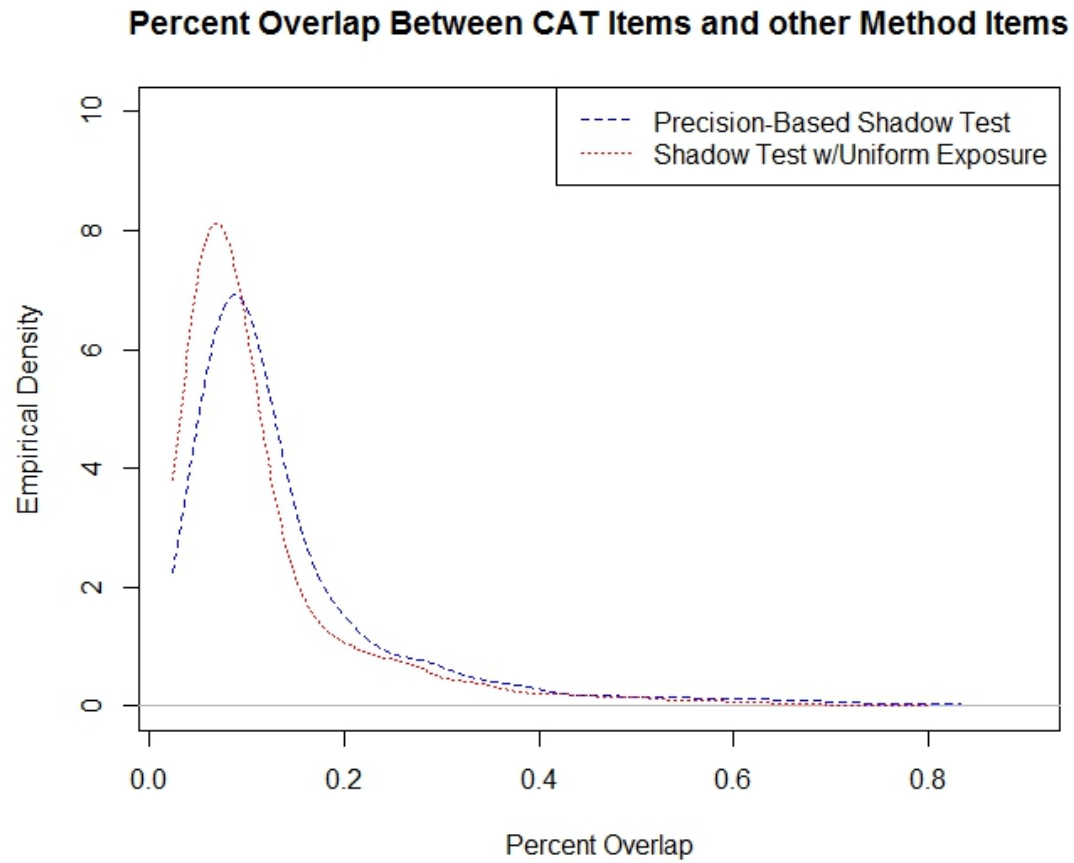


Figure 10. Percentage overlap in item selection by method

Chapter IV: Results

This section will focus on analyses of simulation results pertaining to the quality of ability estimation, item exposure control, and potential for item pilfering (as discussed in earlier sections). These will be compared additionally within and between each of the three item bank size conditions. For the sake of brevity, what we previously called “shadow test with uniform exposure” will be referred to as “UE-ST”, for uniform exposure shadow test (as stated previously, it adapts with respect to ability, and ensures uniform item exposure). “Precision-based shadow test” will now be referred to as “PB-ST” (it adapts with respect to ability and the standard error of ability, and attempts to create uniform exposure). In circumstances where “Nonadaptive” is not a reasonable title for a graph on an aesthetic or formatting basis, “No-Adapt” will be used. Finally, the letter k will refer to size of the item bank whenever used in the context of a figure, table, or elsewhere in the text.

To begin, we discuss quality of ability estimation. Performing well on metrics related to this is a necessary condition for a method to be considered viable.

Estimation of Ability and Standard Error of the Ability Measure

Section 1: Bias. Figures 11-13 provide an expected result; the absence of simulated differential item functioning or other similarly undesirable psychometric phenomena means that each of the test administration/item selection methods performed ability estimation in an unbiased manner. One can attempt to use these figures to glean some initial insights into the method performance with respect to standard error of the ability estimates, but that will be more clearly illustrated in figures 14-16, 20-25, and tables 6-8.

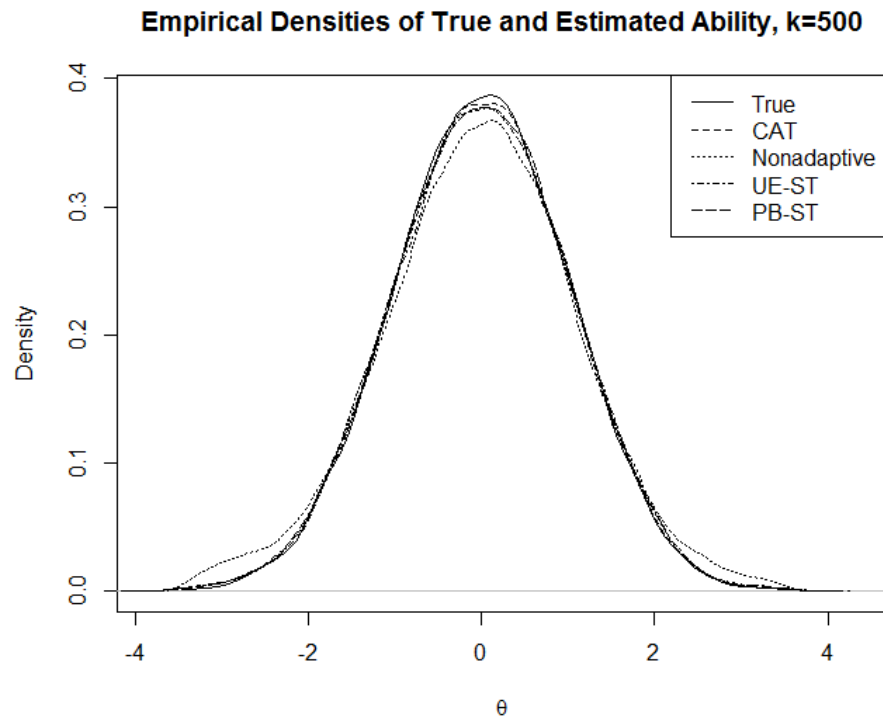


Figure 11. Empirical densities of true and estimated ability ($k=500$)

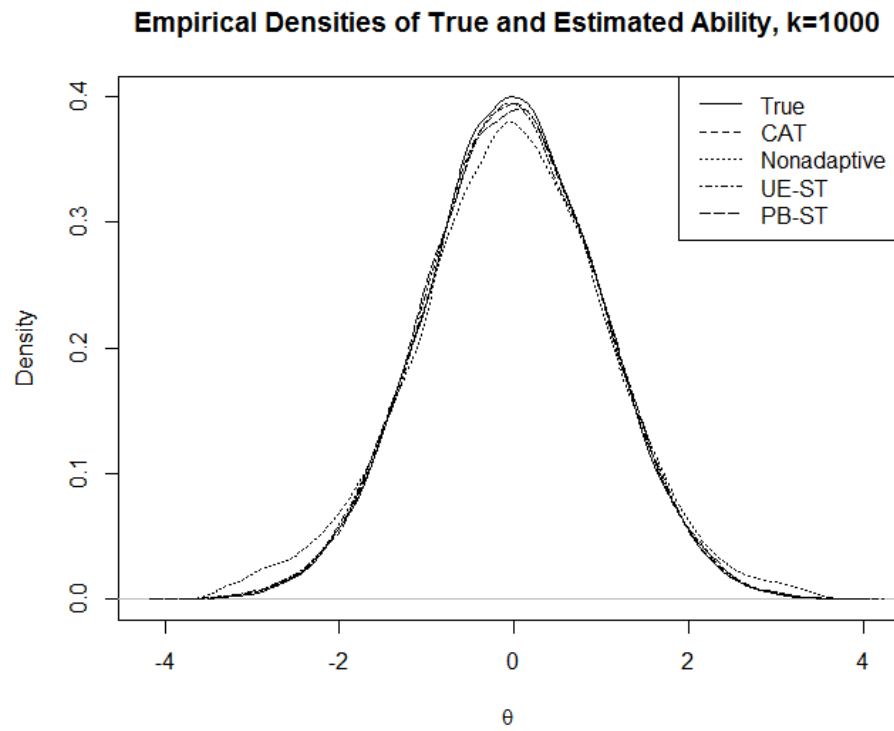


Figure 12. Empirical densities of true and estimated ability ($k=1,000$)

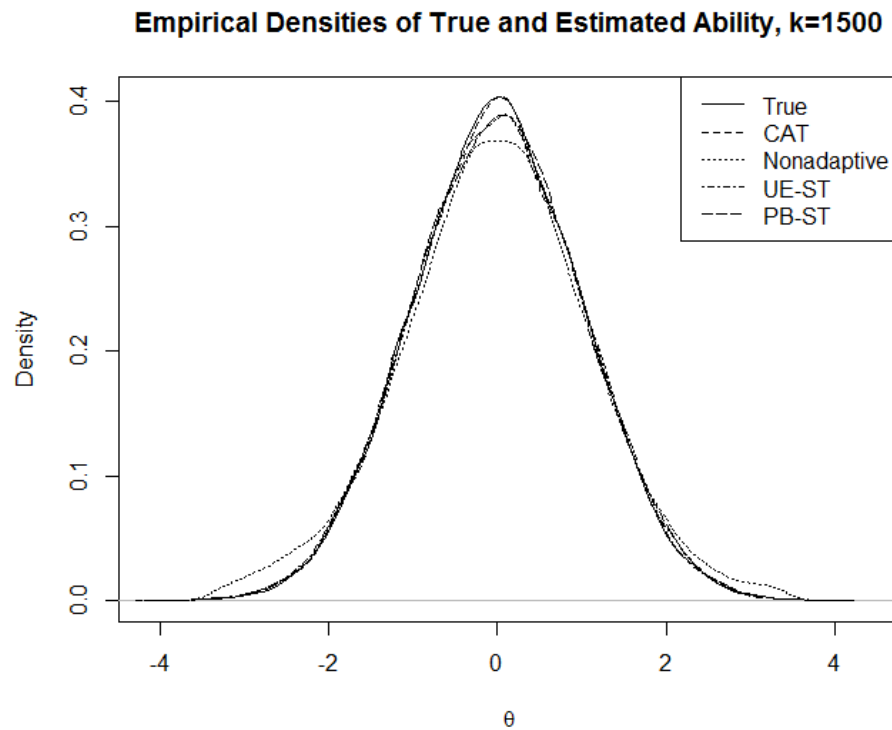


Figure 13. Empirical densities of true and estimated ability ($k=1,500$)

Section 2: Standard Error of the Ability Measure

Part one: Density plots of final standard error estimates. Of particular interest are the final estimates for standard error of the ability estimate -- as seen in figures 14-17. One revealing feature of these results is that the CAT and PB-ST methods exhibit less variability in the standard error estimates than the other methods. In the case of CAT, this is simply because the absence of item exposure control considerations ensures that even the most unusual subjects will still consistently receive informative items, and as such one might characterize this comparatively visually “peaked” distribution (relative to UE-ST and Nonadaptive) as having arisen essentially as a side effect of the estimation procedure.

This differs in the case of PB-ST because precision was used as a component of the item selection process, and as such the tighter control of the final values are deliberate and a direct result of the slight modifications in integer programs differentiating UE-ST and PB-ST.

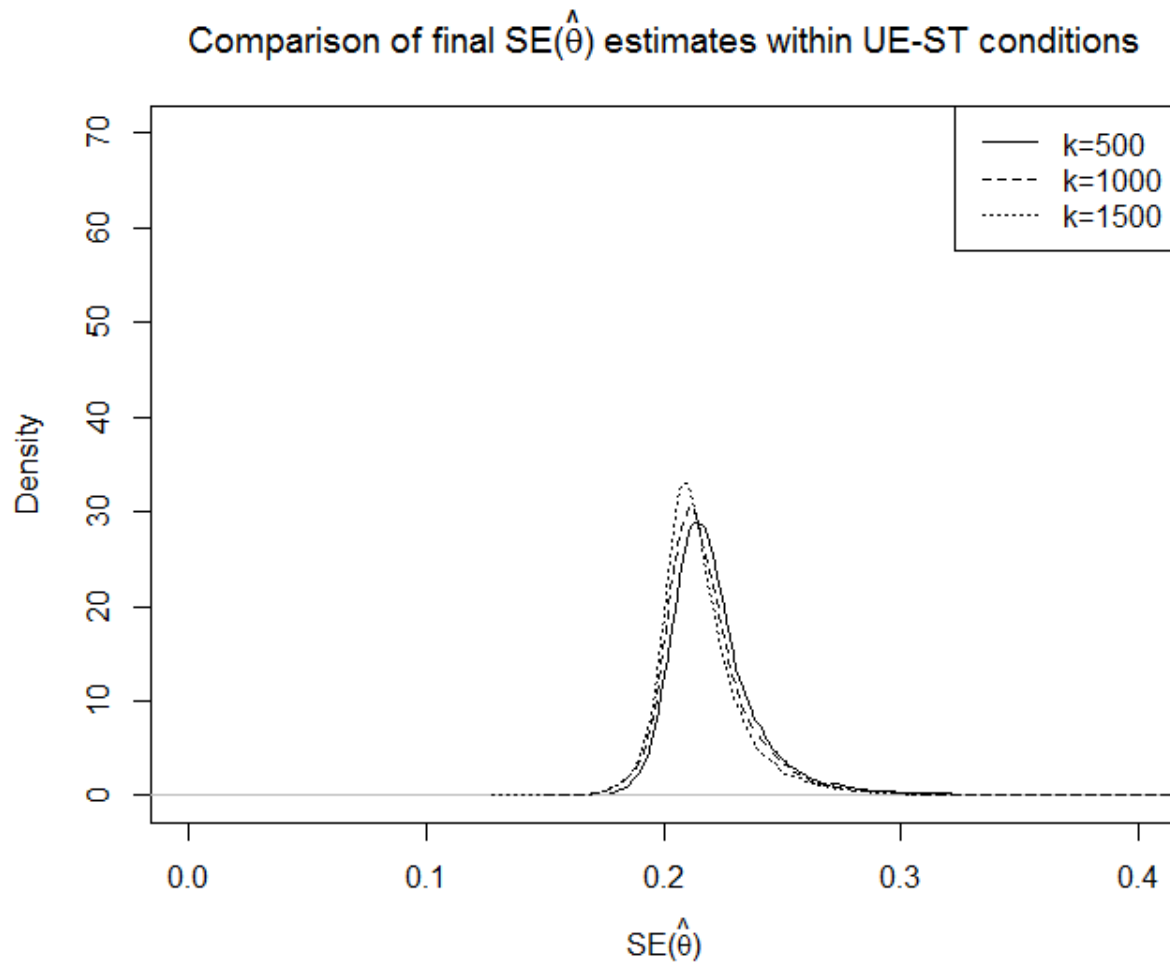


Figure 14. Comparison of final estimates for standard error of the ability estimates within UE-ST Conditions

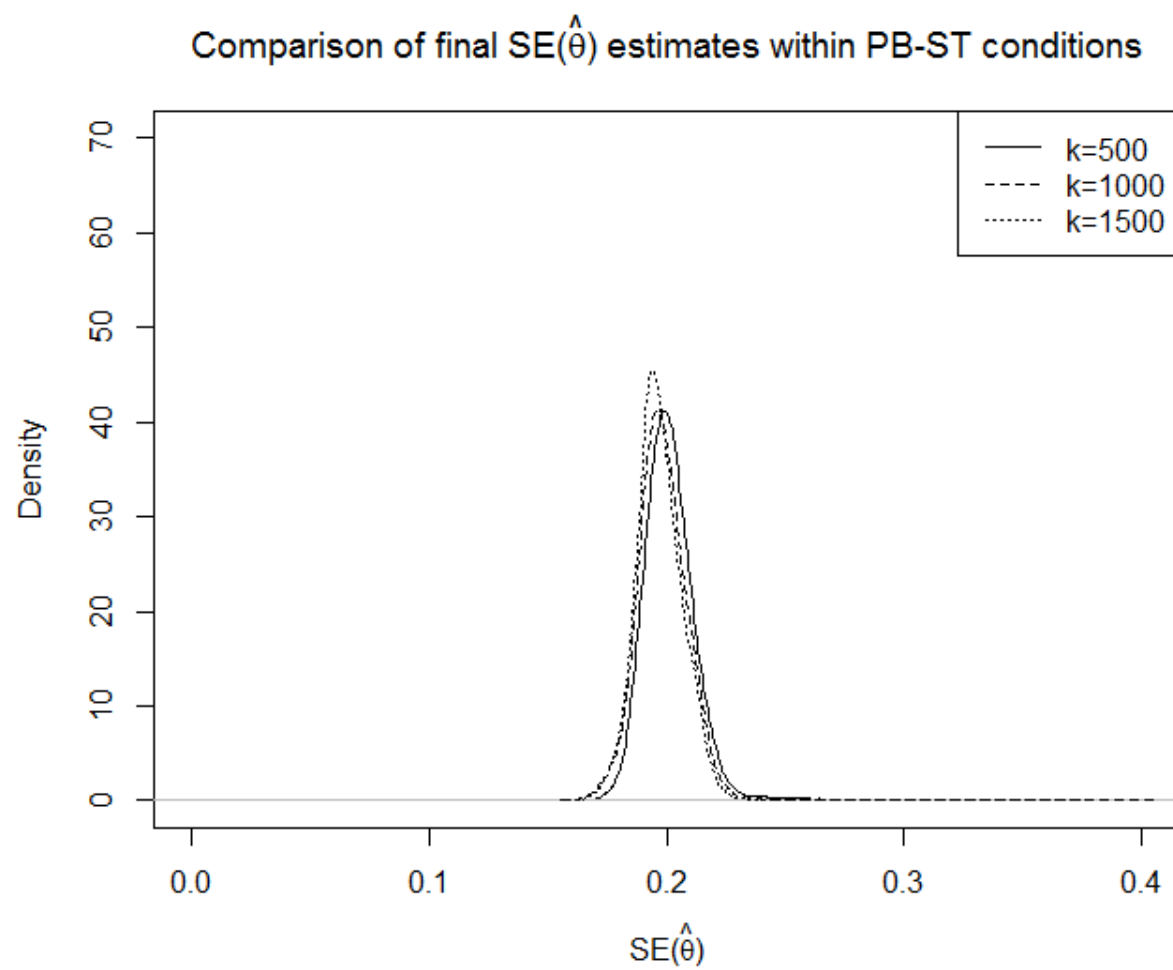


Figure 15. Comparison of final estimates for standard error of the ability estimates within PB-ST conditionss

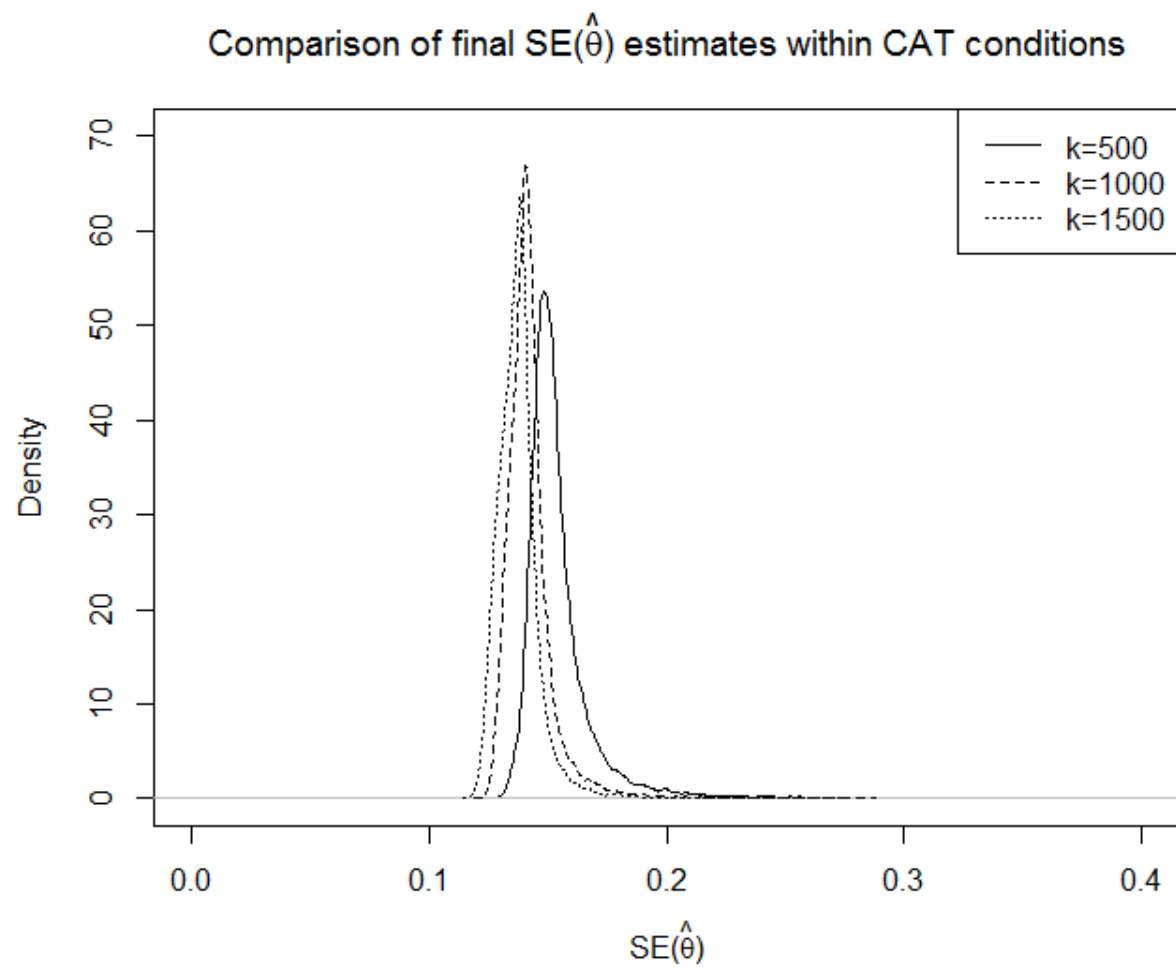


Figure 16. Comparison of final estimates for standard error of the ability estimate within CAT conditions

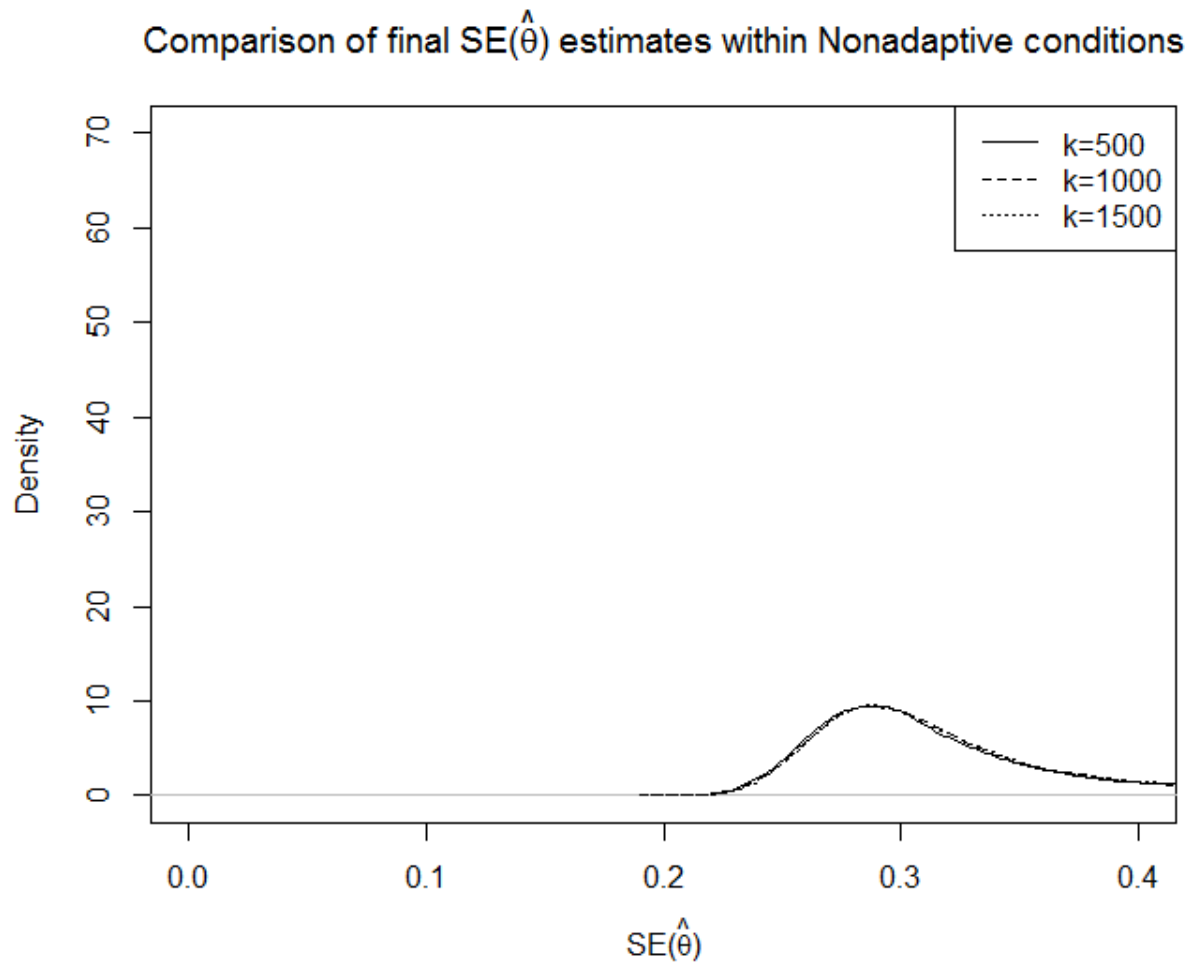


Figure 17. Comparison of final estimates for standard error of the ability estimate within Nonadaptive conditions

Part two: Average $SE(\theta)$ trajectories by method. Figures 18-20 present averages of SE trajectories from the beginning to end of the simulated examinations. Notable in these figures is that CAT relatively unambiguously outperforms all other item selection methods. PB-ST slightly and consistently outperforms UE-ST, both of which substantially outperform Nonadaptive.

There are two distinct classes of trajectory present: in CAT and Nonadaptive, there is no visual evidence that the ongoing improvements in reducing the standard error of the ability estimate would meet diminishing returns if test length was extended. For UE-ST and PB-ST,

however, there is a leveling off of sorts: essentially a horizontal asymptote where the final questions are more devoted to ensuring uniform item exposure than increasing precision of the ability estimate for all subjects. Extrapolation based upon this leveling off is misguided, though; it is not the case that if test length were changed that the methods would fail to improve precision beyond the currently 40-item test, as the formulation of these methods would adjust the maximum exposure rates for items to compensate for this change.

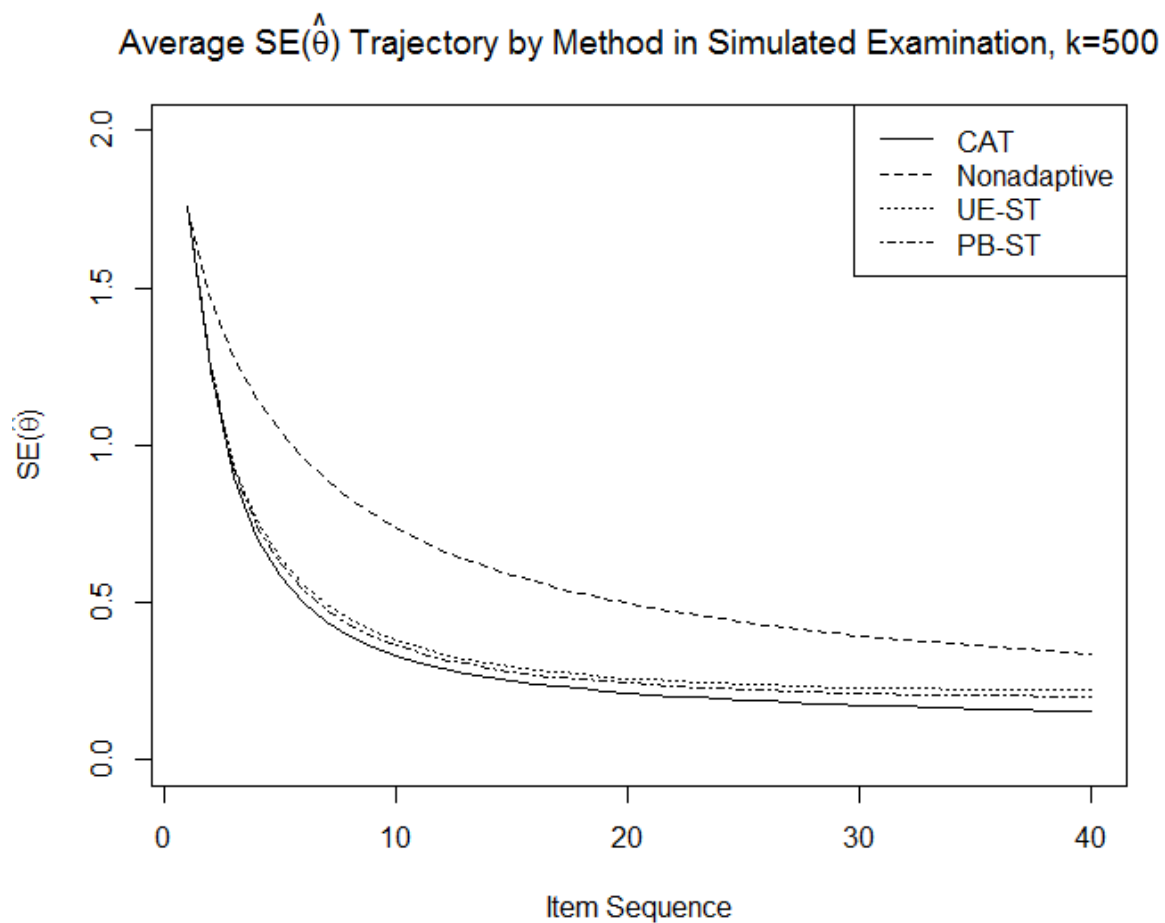


Figure 18. Averages of $SE(\theta)$ trajectory by method in simulated examination ($k=500$)

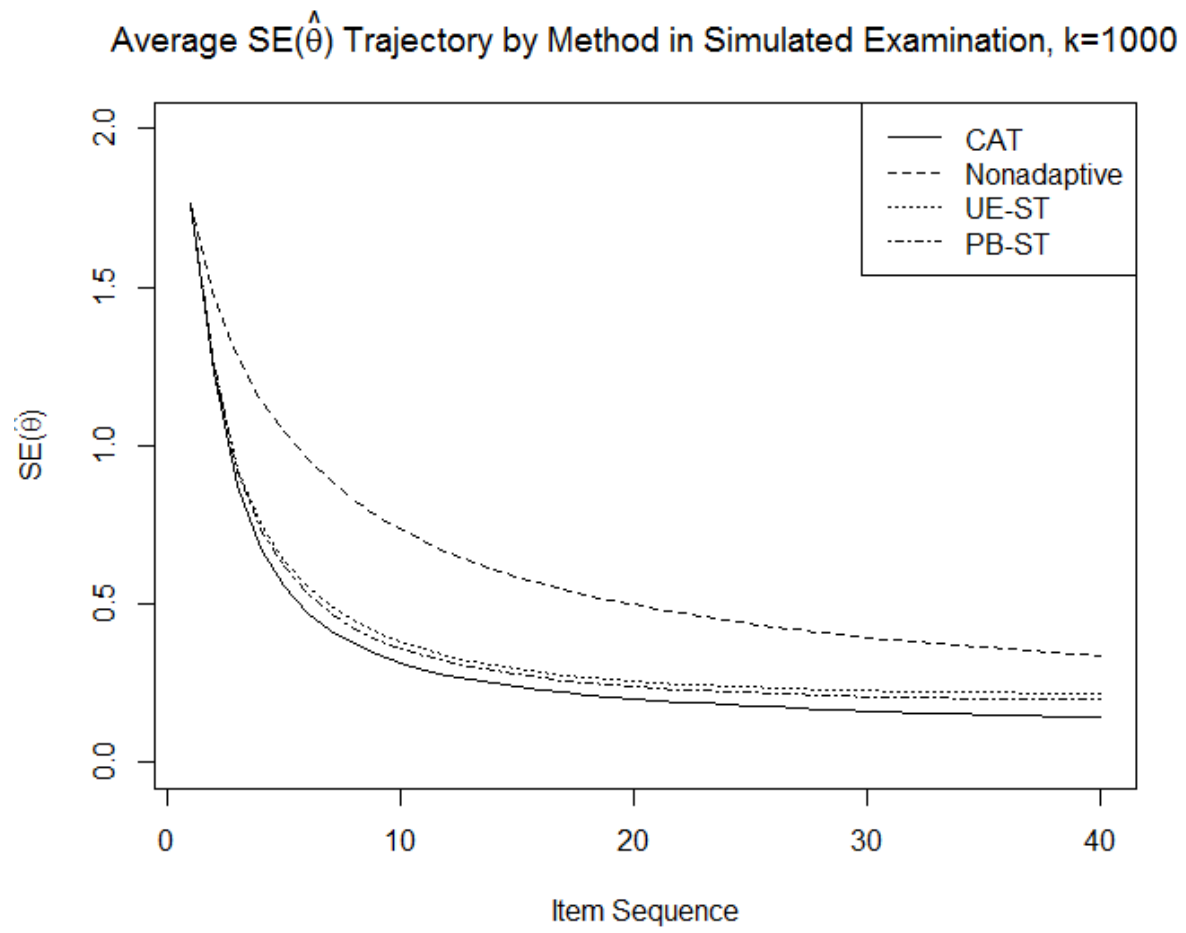


Figure 19. Averages of $SE(\theta)$ trajectory by method in simulated examination ($k=1,000$)

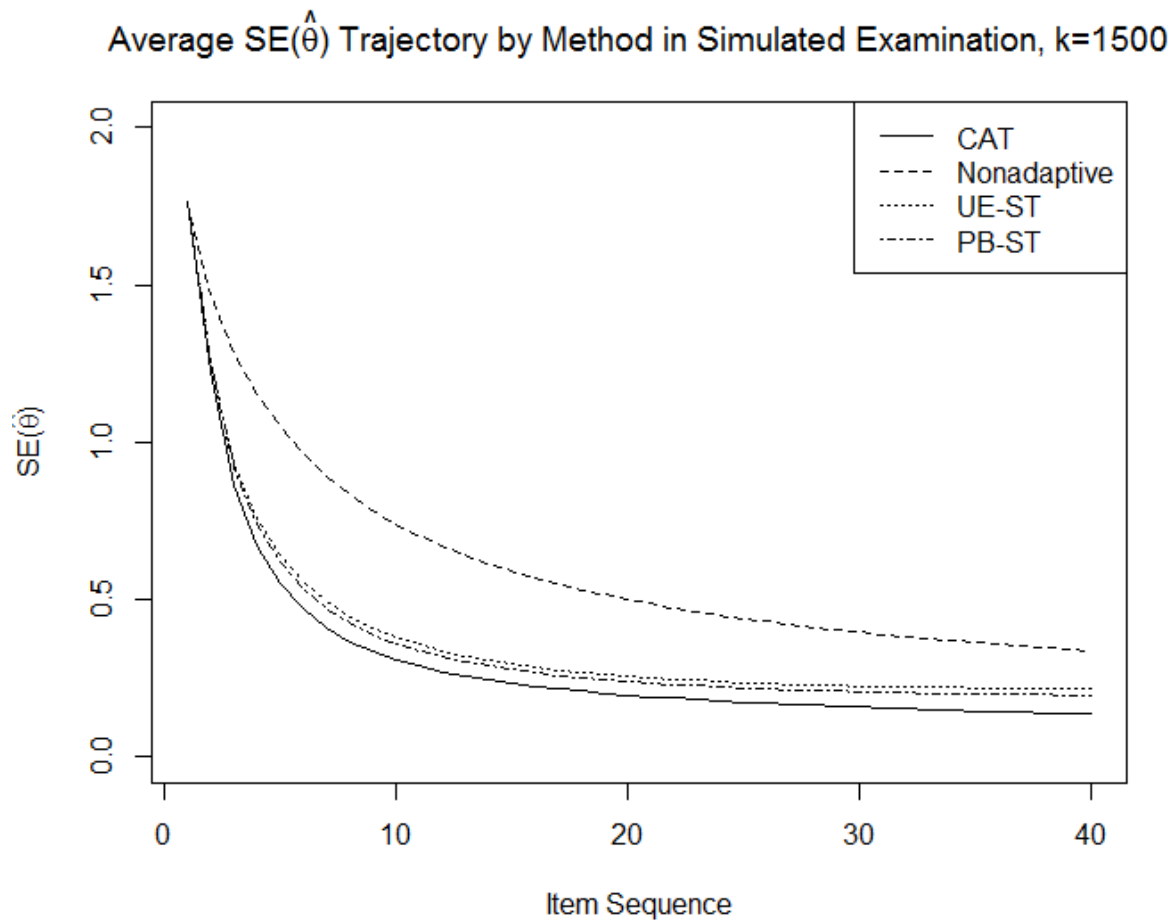


Figure 20. Averages of $SE(\theta)$ trajectory by method in simulated examination ($k=1,500$)

Part three: Relationship between θ and $SE(\theta)$. The plots featured in figures 21-24 are comprised of the same exact standard error estimates from figures 14-17, with the addition of true ability for the subject along the x-axis. These figures are informative for a few reasons.

First, the clearly parabolic nature of the $SE(\theta)$ by θ observations in the CAT and Nonadaptive methods is contrasted with the obviously flat, linear shapes of the $SE(\theta)$ by θ observations in UE-ST and PB-ST. In the case of CAT, it is not surprising that subjects with ability closest to the average of the item bank's b parameters have greatest precision in their ability estimate. Similarly for Nonadaptive, where one would expect such a shape to arise

naturally, mimicking the distribution of the b parameters in the item bank. The linearity (at least from approximately $[-2,2]$ on the latent continuum in the $k=500$ case, and expanding further from there as item bank size increases) is a desirable psychometric property in the sense that ability and SE of ability are unrelated. This is a rare but ideal case in application, in which one can honestly state that all subjects are expected to finish with approximately equal “confidence” in their ability estimates, regardless of what that ability may be.

In the item bank size of 500 condition, the region of the latent continuum outside of the previously-mentioned $[-2,2]$ represents subjects that do not receive similarly high-precision estimates of ability, regardless of item selection method. This is, in effect, a barometer of sorts where one may state that there does not exist sufficient information in the item bank to reach a high-precision estimate for particularly high and particularly low ability subjects. As this range expands in the $k=1,000$ and $1,500$ conditions, one may reasonably conclude that both CAT and PB-ST are effectively using the larger item pool to improve ability estimation for unusual levels of ability.

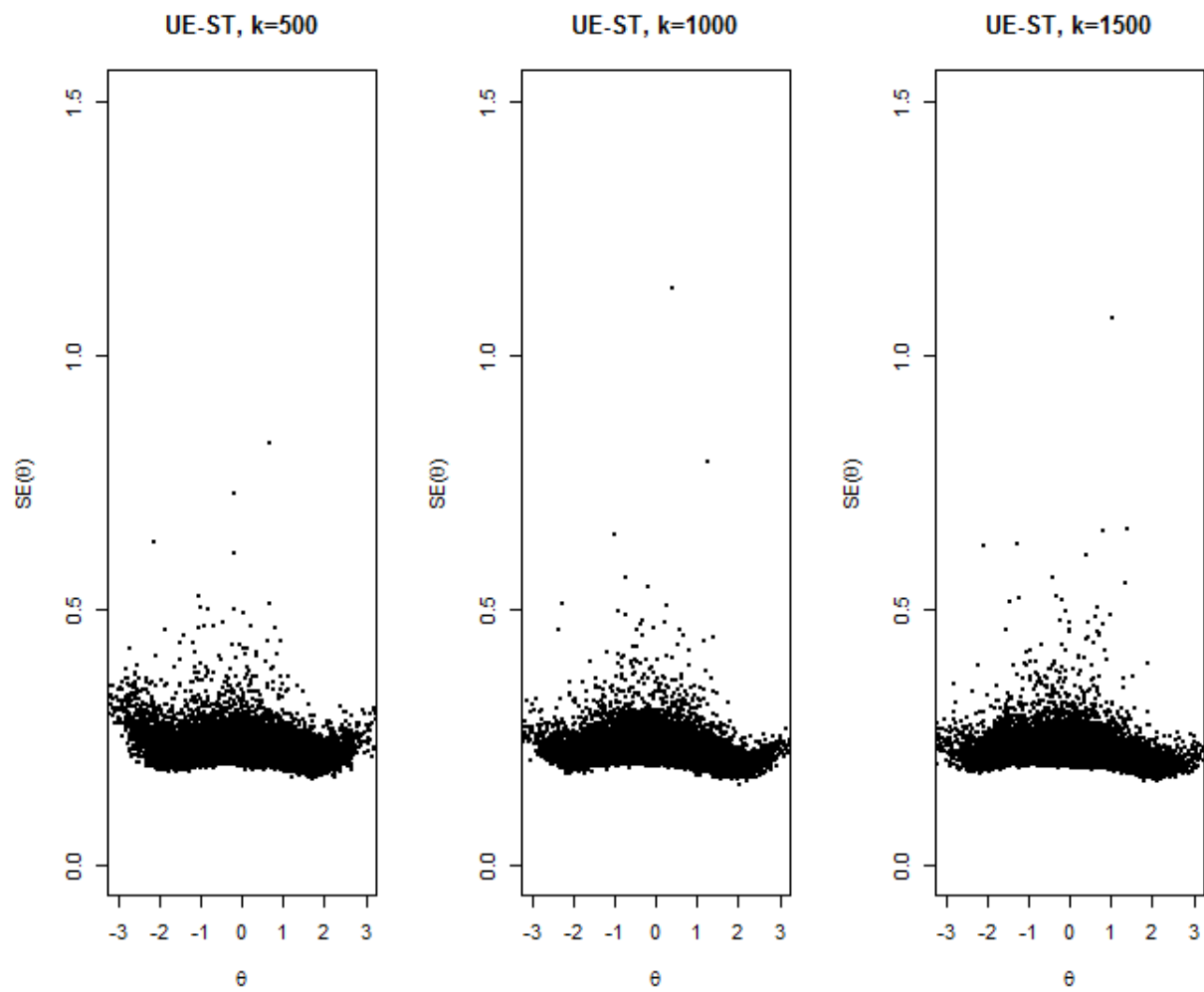


Figure 21. Relationship between true ability and standard error estimates within UE-ST conditions

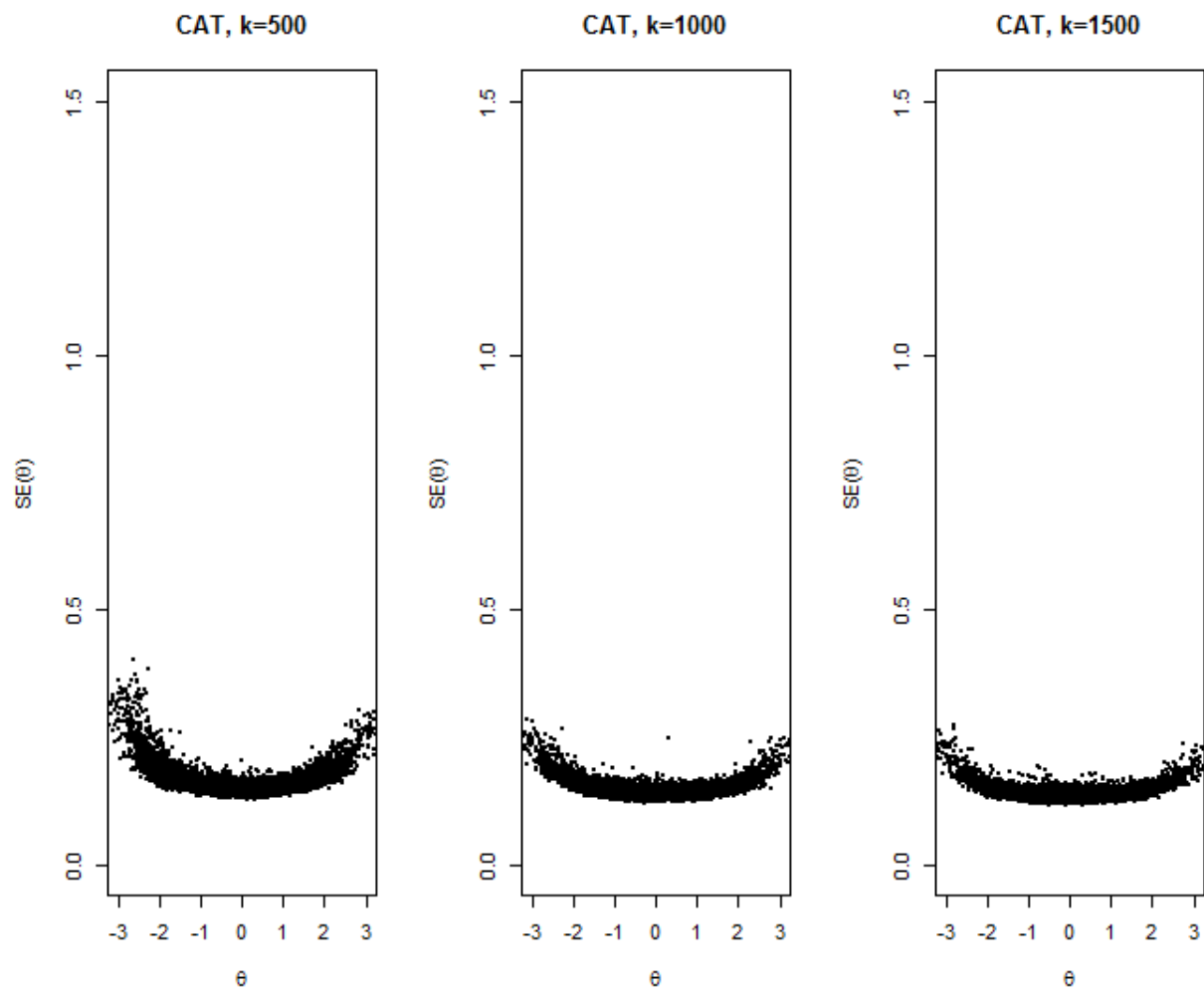


Figure 22. Relationship between true ability and standard error estimates within CAT conditions

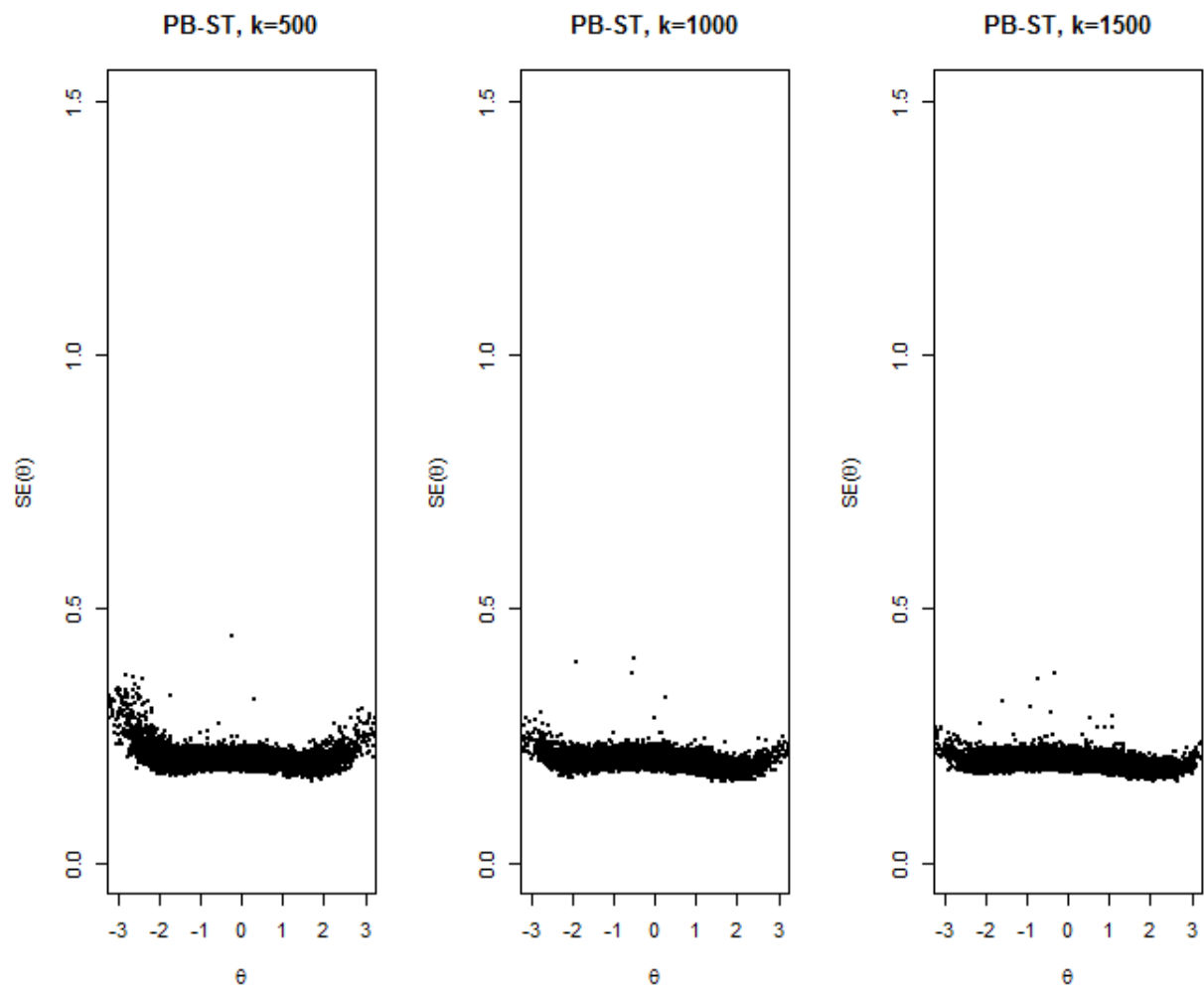


Figure 23. Relationship between true ability and standard error estimates within PB-ST conditions

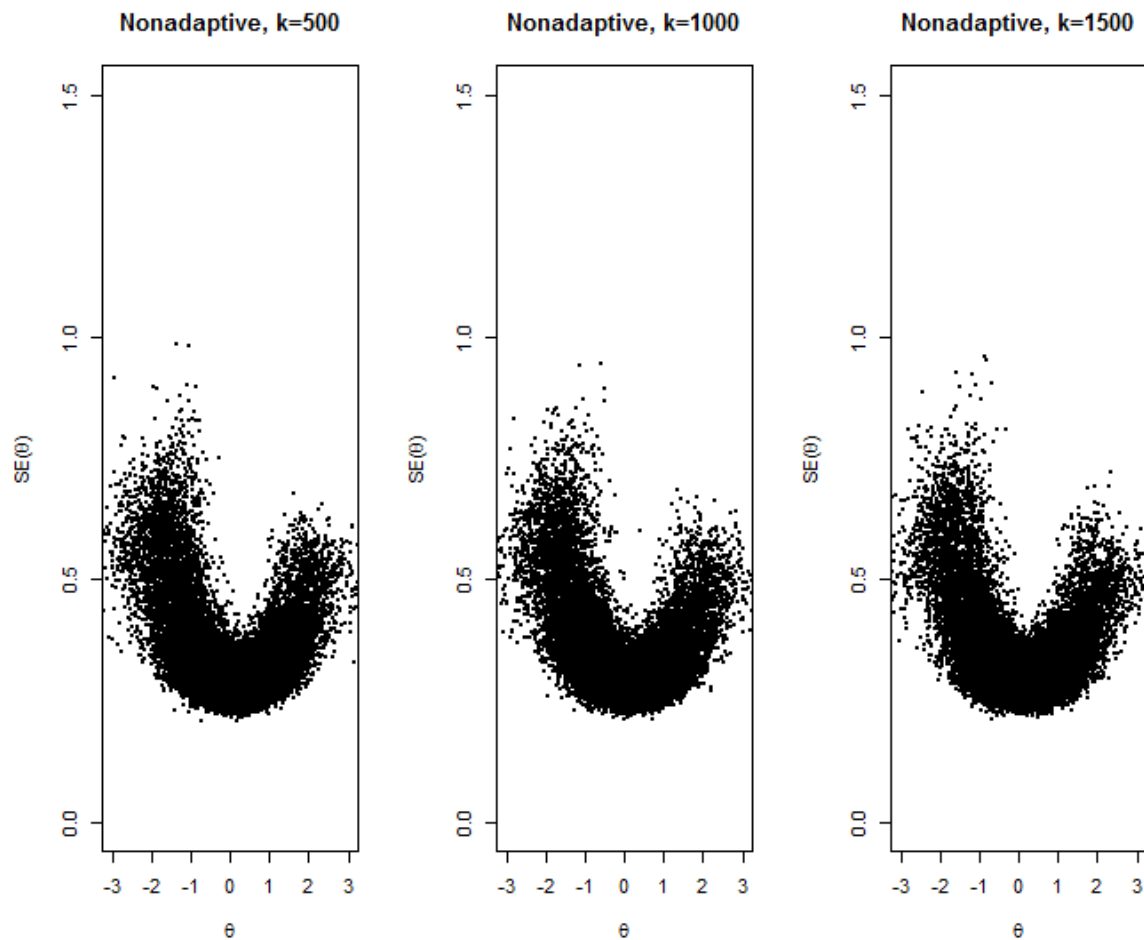


Figure 24. Relationship between true ability and standard error estimates within Nonadaptive conditions

On the other hand, UE-ST appears to at best not improve, at worst worsen with regard to controlling the distribution of $SE(\theta)$. This is a finding that discussed in-depth after the introduction of a few figures later in this section; the combined information from the figures (and accompanying tables) reveal the causes of this dysfunction.

Finally, it is certainly visually apparent that the left side of the Nonadaptive parabola extends higher than the right, and to a much lesser extent this is present in each method. The reason for this is because this simulation was conducted under a 3PL IRT model, and the guessing parameter is disproportionately punitive to standard error of ability estimates for low-ability subjects. This is elaborated upon later in this section as well.

Part four: other properties of $SE(\theta)$ estimates. A natural follow-up to the θ by $SE(\theta)$ scatterplots are additional analyses investigating the differences in distribution of $SE(\theta)$ by method. Figures 25-27 each provide three boxplots representing final $SE(\theta)$ distributions for the featured item selection methods, separated by item bank size. The information these figures provide beyond what is already provided in the layered density plots from figures 14-17 is the inter-quartile ranges, outlier classification and easy sight approximation of range. In other words, a key piece of information obfuscated in the scatterplots is density of observations and a key piece of information obfuscated in kernel density plots is how meaningful visual differences in distributional shape translate to the actual values contained within the distributions (e.g. one “tail” of a distribution could look like it contains no observations when in fact there are hundreds): hence, the need for figures 25-27.

Here, as before, CAT and PB-ST demonstrate superior control of homogeneity in $SE(\theta)$ than the other two methods, as well as in the prevention of more severe outliers. Interestingly, as supported by tables 6-8 (whose originating data created figures 25-27 here as well as figures 14-16), no method aside from CAT improves its average $SE(\theta)$ as item bank size increases. It is also one of two methods that improves on other measures (e.g. range). The reasons for this differ by method.

Table 6. Summary statistics of final $SE(\theta)$ estimates by method ($k=500$)

Method, $k=500$	N	Mean	Median	SD	Min	Max	Range	Skewness	Kurtosis
<i>CAT</i>	26,250	0.15	0.15	0.02	0.13	0.4	0.27	4.59	31.95
<i>Nonadaptive</i>	26,250	0.34	0.31	0.09	0.21	0.99	0.78	2.05	4.75
<i>UE-ST</i>	26,250	0.22	0.21	0.02	0.17	0.83	0.66	4.18	47.79
<i>PB-ST</i>	26,250	0.20	0.20	0.01	0.16	0.44	0.28	3.30	27.11

Table 7. Summary statistics of final $SE(\theta)$ estimates by method ($k=1,000$)

Method, k=1,000	N	Mean	Median	SD	Min	Max	Range	Skewness	Kurtosis
<i>CAT</i>	26,250	0.14	0.14	0.01	0.12	0.29	0.17	3.75	24.76
<i>Nonadaptive</i>	26,250	0.34	0.31	0.09	0.21	0.94	0.73	2.08	4.96
<i>UE-ST</i>	26,250	0.22	0.21	0.02	0.16	1.13	0.98	5.90	130.53
<i>PB-ST</i>	26,250	0.20	0.20	0.01	0.16	0.40	0.24	1.12	13.92

Table 8. Summary statistics of final $SE(\theta)$ estimates by method ($k=1,500$)

Method, k=1,500	N	Mean	Median	SD	Min	Max	Range	Skewness	Kurtosis
<i>CAT</i>	26,250	0.14	0.14	0.01	0.12	0.27	0.15	3.39	25.06
<i>Nonadaptive</i>	26,250	0.34	0.31	0.09	0.21	0.96	0.75	2.04	4.74
<i>UE-ST</i>	26,250	0.22	0.21	0.02	0.13	1.07	0.94	6.30	119.47
<i>PB-ST</i>	26,250	0.20	0.20	0.01	0.16	0.37	0.22	0.96	9.54

In this case, Nonadaptive does not change on *any* of the distributional characteristics across the item bank size conditions, which is precisely what one might expect. The “new” items have their psychometric properties drawn from the same exact distributions as the “old” items, so the approximately parallel forms created via non-adaptive item selection will only change with respect to the number of possible forms and not with respect to the distribution of test information represented by the forms.

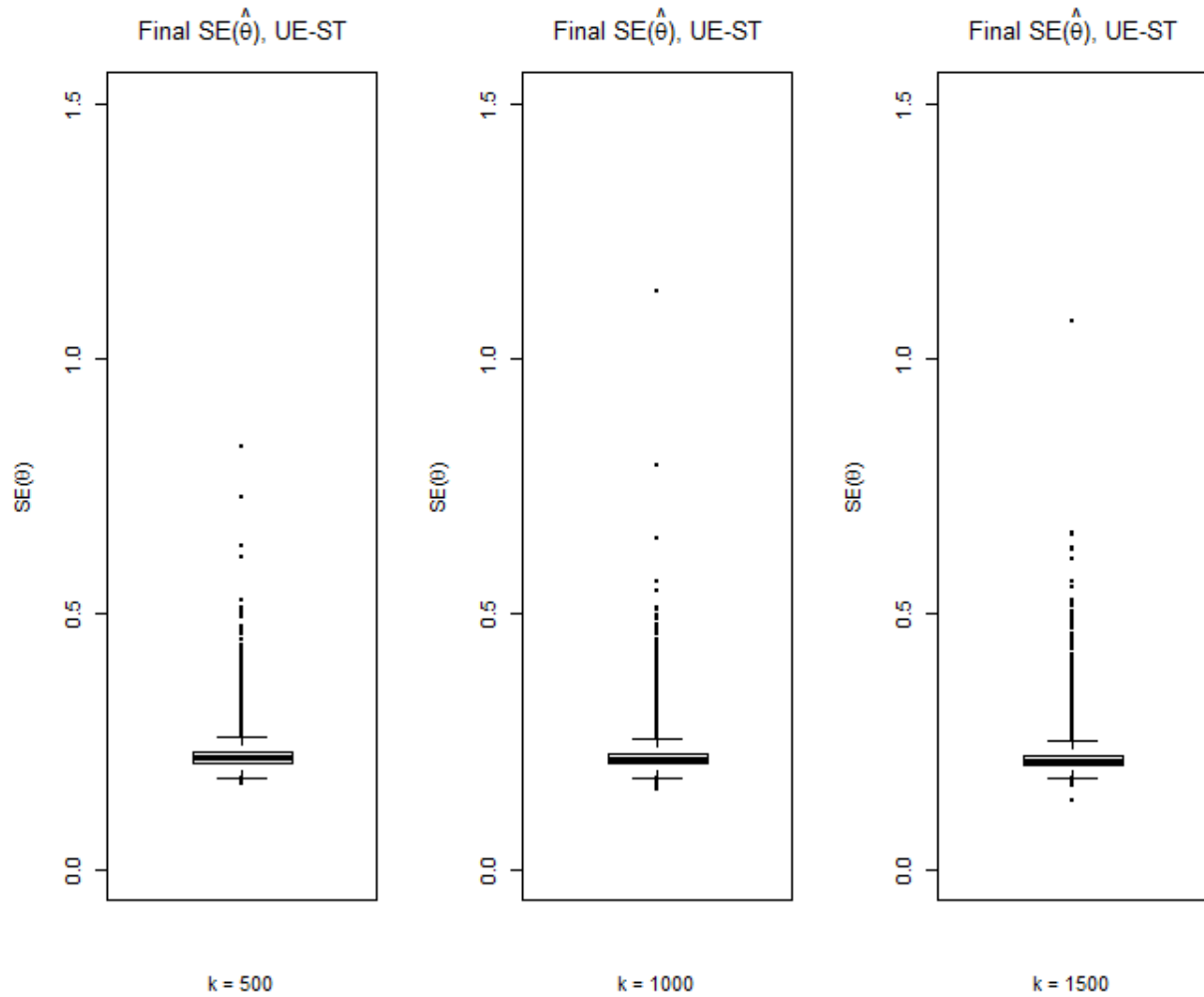


Figure 25. Box-plots of $SE(\theta)$ within UE-ST conditions

UE-ST, on the other hand, undergoes a highly instructive fall from grace; relative to the item bank size of 500 condition, the 1,000 and 1,500 conditions have higher maxima, larger ranges, and higher skewness (though the maxima and ranges are partially driven by slightly improved minima). These are all the result of the same phenomenon inherent to how such a test is administered: items are given out on a “first come, first served” basis. This means that subjects who navigate faster to their true ability region are rewarded with the most informative items from that region, and it is the driving factor behind the lowering minimum standard error observed. The downsides of such a strategy can be described in an all-too-common scenario: a

student performs inconsistently at the start of an examination, but ultimately starts getting items with appropriate difficulty. The problem is that other people with similar ability have already gotten there and taken the highly discriminating, low guessing, “good” items. In CAT this is not a problem because items one subject receives are independent of items another receives. In PB-ST, this is not a problem because the precision/“need”-based item selection will simply administer good items until this subject has “caught up” -- it was designed to handle the negative consequences of unusual or volatile response patterns. In Nonadaptive, such a subject would still get the same questions regardless of their response pattern and previous questions, so the temporal component driving the high $SE(\theta)$ does not exist; questions are equally influential in ability estimation regardless of when along the test sequence they are administered. But all of this is simply the reason why there are so many high $SE(\theta)$ in the middle of the θ continuum; how does item bank size exacerbate the problem?

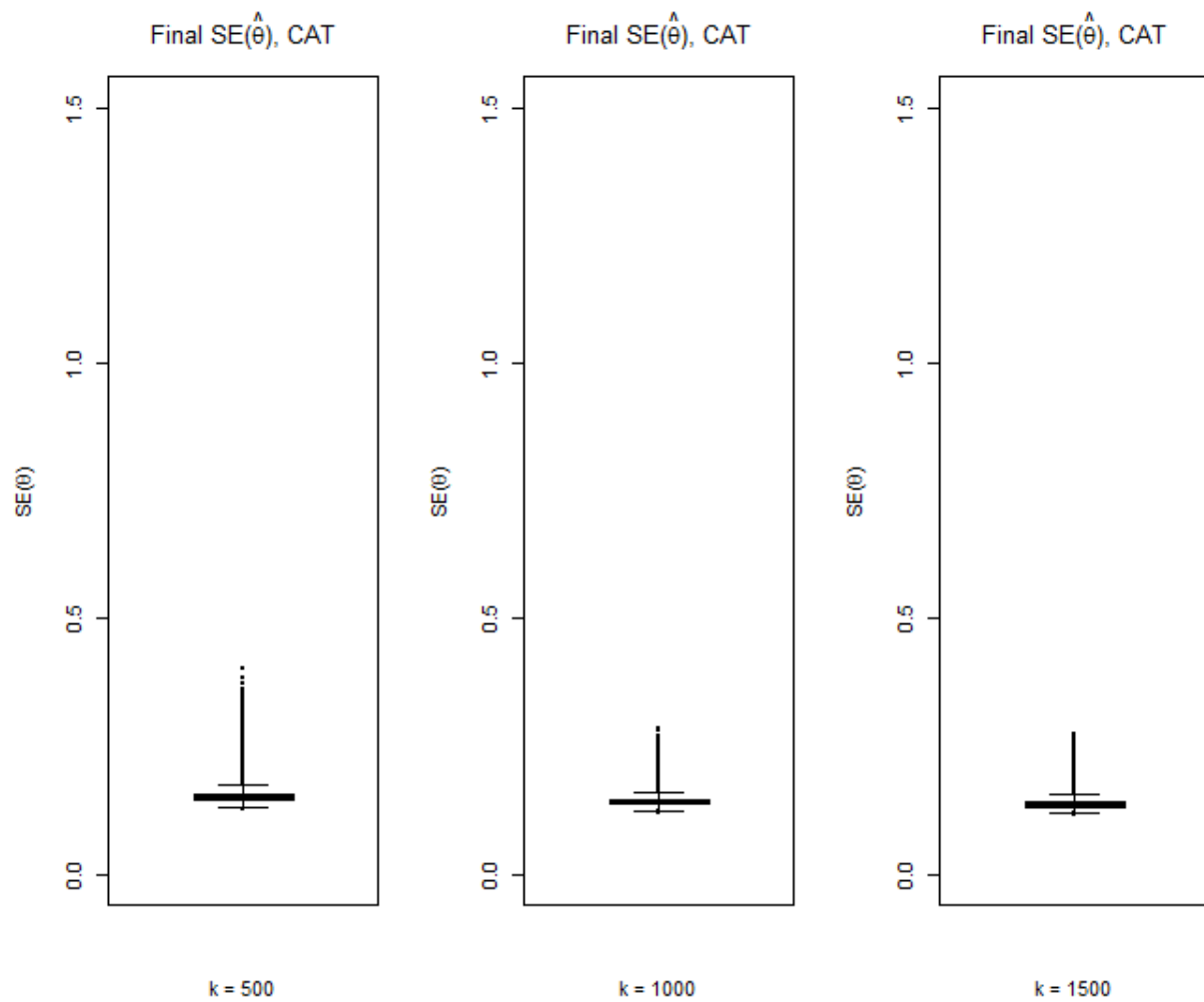


Figure 26. Box-plots of $SE(\theta)$ within CAT conditions

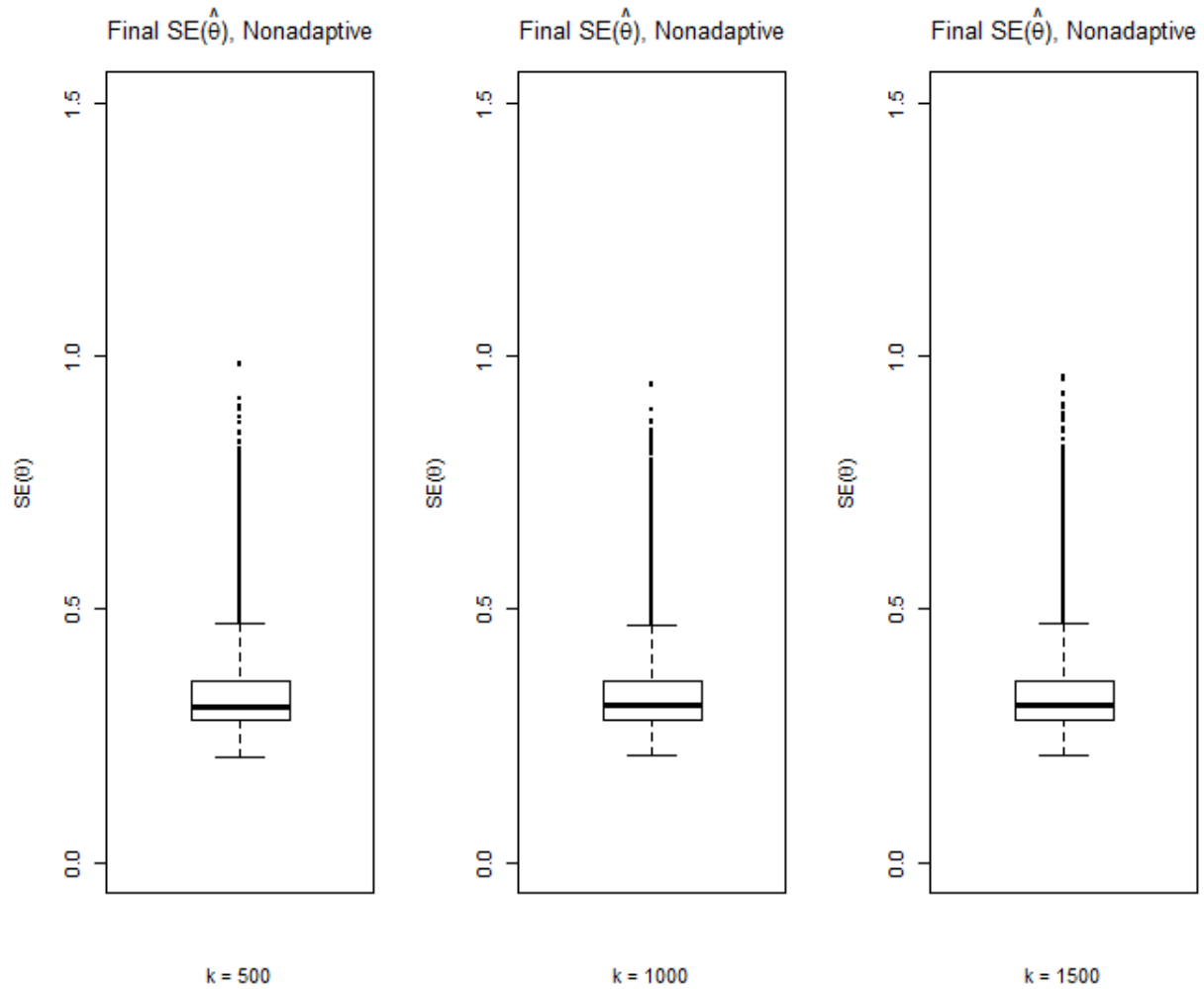


Figure 27. Box-plots of $SE(\theta)$ within Nonadaptive conditions

The problem is exacerbated because the test length and number of examinees remains the same. As such, the value for frequency of item selection actually lowers as item bank size increases. More specifically, for the 500, 1,000 and 1,500 conditions, the frequencies indicating uniform exposure were 60, 30, and 20, respectively. It appears to be the case that although there are more items to choose from, the exactly equal proportional lowering of maximum permitted exposure causes more harm than more items to choose from does good. The reason for this is the following: previously, up to 60 examinees could take a single item. This means that late arrivals, so to speak, still have a reasonable chance at seeing a high-quality item, given that there

are only 750 subjects and no one is allowed to see an item more than once. On the contrary, when only 20 subjects are allowed to see an item, the “early birds” arrive to a veritable buffet of high-quality items, all of which they could push to maximum exposure before the late arrivals. It is a worse scenario for the late arrivals, and a better one for the early birds, which is why the minimum $SE(\theta)$ went down a little and the maximum $SE(\theta)$ soared.

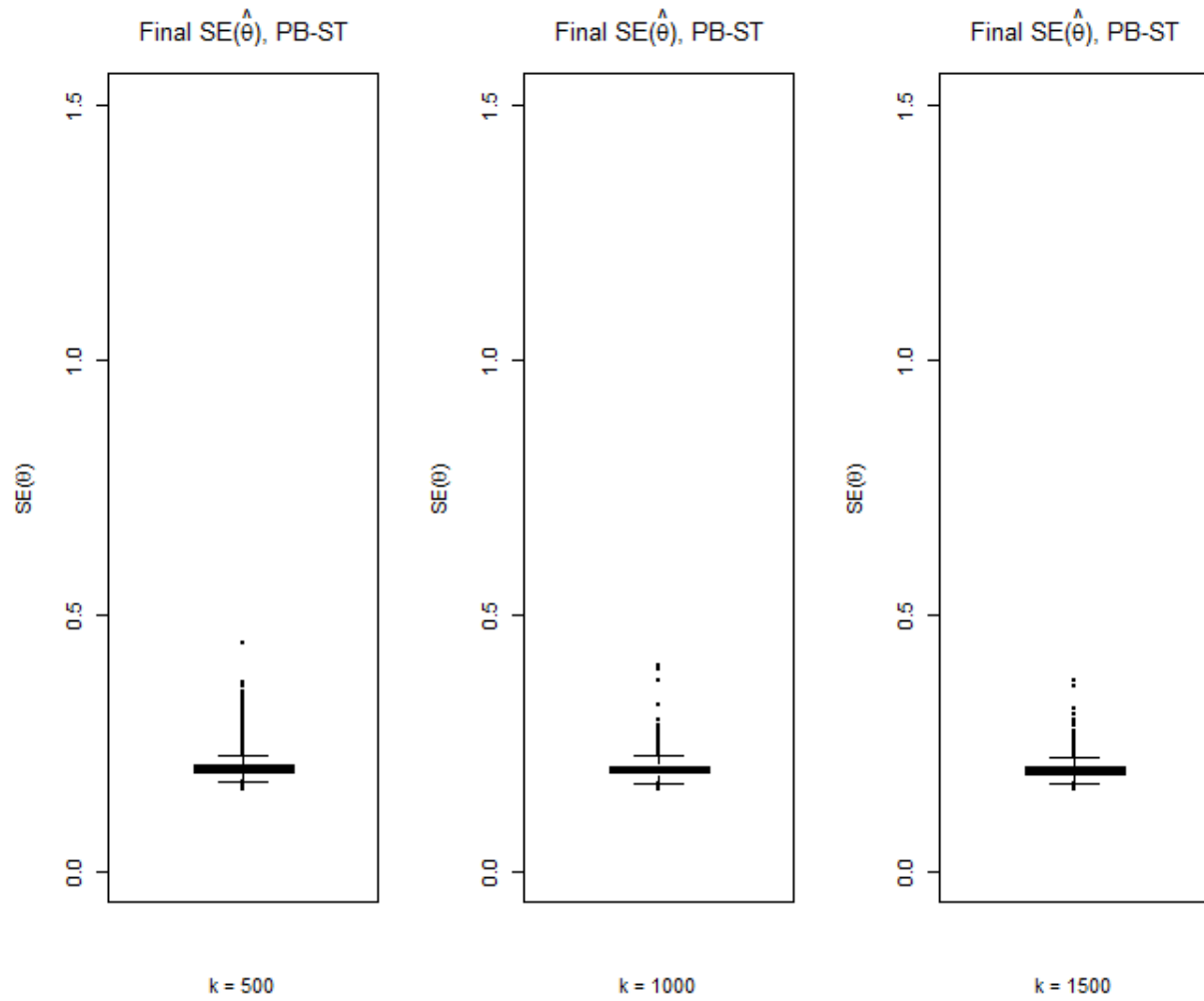


Figure 28. Box-plots of $SE(\theta)$ within PB-ST conditions

Rounding out this discussion is the similarly interesting performance of PB-ST as item bank size increased: maximum $SE(\theta)$ decreased, range decreased, skewness decreased, and

kurtosis decreased. The method administers suboptimal items to subjects with lower-than-expected $SE(\theta)$ at all junctures during the exam administration, so the potential gain from having a deeper item pool is deliberately not manifested in improvement in minimum $SE(\theta)$. The improvements are directed specifically at subjects with less precise estimates of ability, as they are the ones whose “need” score will justify the administration of the most informative items. By tactically administering low-information items when it is not harmful and high-information items when it is necessary, the distribution of $SE(\theta)$ tightens and becomes nearly symmetric in the item bank size of 1,500 condition. This, in conjunction with the earlier scatterplots demonstrating this homogeneity in $SE(\theta)$ across the latent continuum, is highly desirable from a psychometric and practical perspective. It is psychometrically desirable because the elimination of outliers and other fringe cases suggests a more reliable testing process, with the added bonus of (with a sufficiently large item pool) a symmetric distribution of $SE(\theta)$. Speaking more to the pragmatic concerns inherent to large-scale, high-stakes ability testing, it is likely desirable to be able to honestly state that estimates of examinee ability are equally precise regardless of where on the ability continuum and to simultaneously address the very different problems of volatile early performance and item pilfering using the same statistical mechanism.

Part five: additional distributional comparisons of $SE(\theta)$ by method. Figures 29-37 are Q-Q plots where $SE(\theta)$ estimate distributions are compared by method, with an identity line superimposed for reference. These are useful in determining at precisely what values and to what degree one method outperforms another with respect to the standard error of the ability estimate.

As an example, across the item bank size conditions in figures 29-31 one may observe CAT go from slightly outperforming PB-ST on the lower end of the standard error spectrum to a more thorough and uniform outperformance across the entirety of the standard error spectrum. In the general sense that larger deviations from the superimposed identity line indicate larger numerical differences in the raw numbers that constitute the distributions, we can state that the hierarchy from most to least precise is CAT, PB-ST, UE-ST and then Nonadaptive.

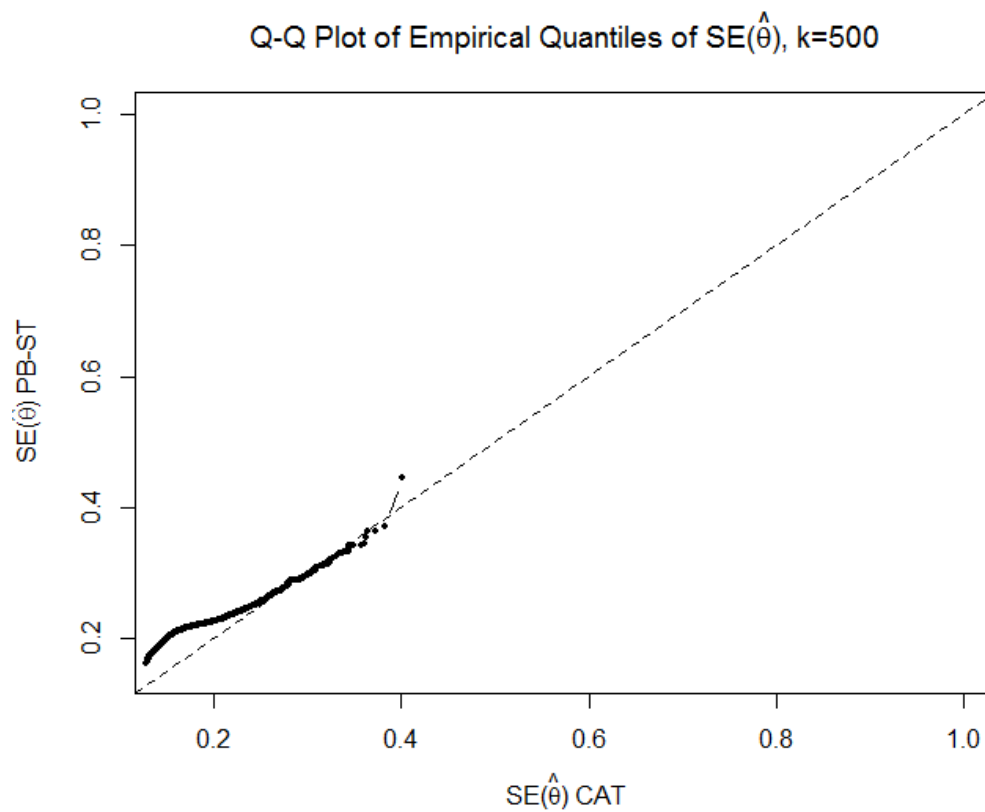


Figure 29. Q-Q plots of $SE(\theta)$ estimate distributions: CAT vs PB-ST ($k=500$)

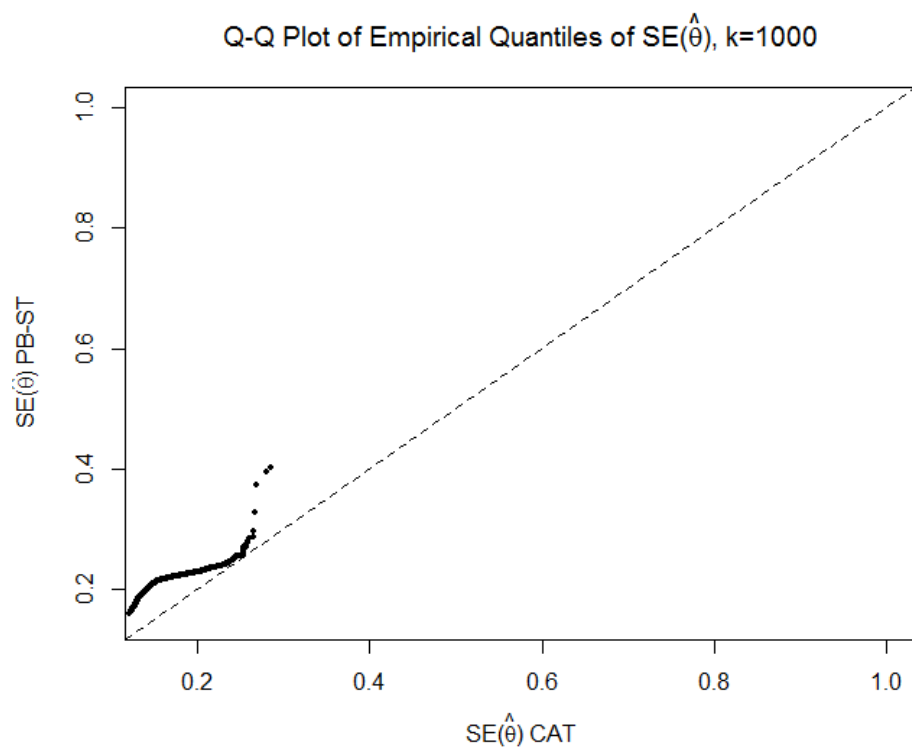


Figure 30. Q-Q plots of $SE(\theta)$ estimate distributions: CAT vs PB-ST ($k=1,000$)

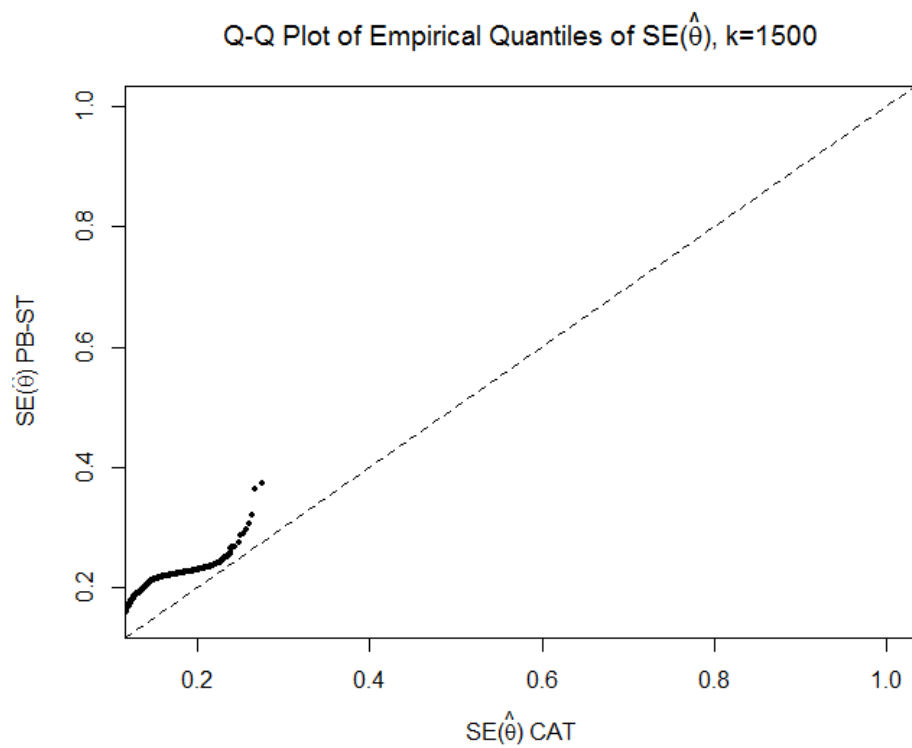


Figure 31. Q-Q plots of $SE(\theta)$ estimate distributions: CAT vs PB-ST ($k=1,500$)

The differences between CAT and PB-ST are exacerbated further as item bank size increases, which is somewhat ironic in the sense that one might be tempted to call it a more efficient exploitation of the introduction of greater item selection possibilities, when the statistical reality is this improvement is achieved through inefficient use of the item bank. As our discussion of item exposure results will later reveal, CAT is not selecting more items to use; the case here is that the ~200 items CAT does use are more likely to have advantageous psychometric qualities when choosing out of 1,500 possible items rather than choosing from 1,000 or 500. Neither the test length nor the sample size changed; CAT in effect constructed the same house, only this time with higher quality materials.

The comparison of PB-ST and Nonadaptive using figures 29-31 is what one might expect when comparing an adaptive against a nonadaptive test. As discussed previously with respect to trajectories of the standard errors across the entirety of the examination, Nonadaptive simply needs more items to achieve equivalent levels of precision: a result which is nearly as reliable as gravity with respect to reproducibility. The exact shape taken by the lines featured in figures 24-26 is the result of Nonadaptive's ability to do well precisely estimating ability for average ability subject but having substantial difficulty with any subjects deviating even slightly from that specific central area of the ability distribution. In administering items non-adaptively, Nonadaptive will provide a test that caters to whatever population is most in accordance with the characteristics of the item bank; unlike CAT, UE-ST and PB-ST, it has no recourse and will perform poorly if there exists incongruity between the population ability distribution and the item bank properties (most importantly, the b parameter).

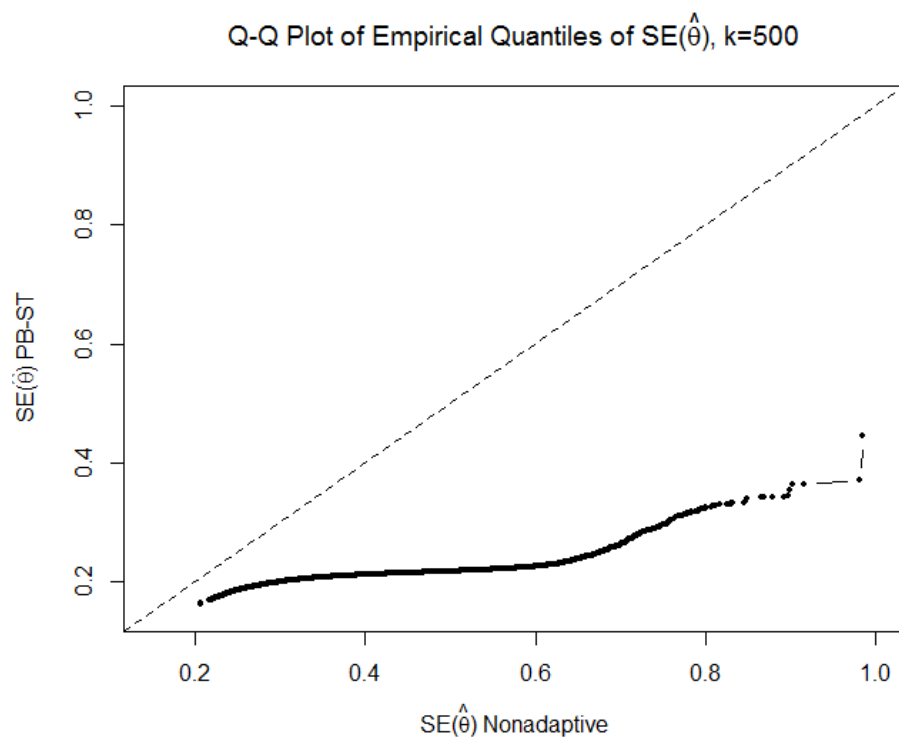


Figure 32. Q-Q plots of $SE(\theta)$ estimate distributions: Nonadaptive vs PB-ST ($k=500$)

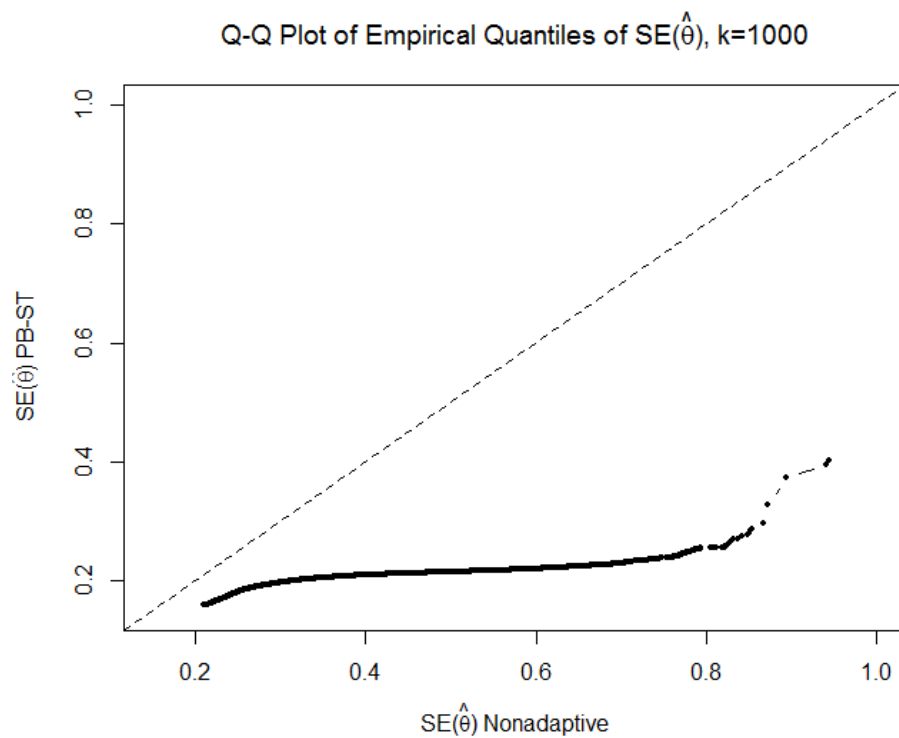


Figure 33. Q-Q plots of $SE(\theta)$ estimate distributions: Nonadaptive vs PB-ST ($k=1,000$)

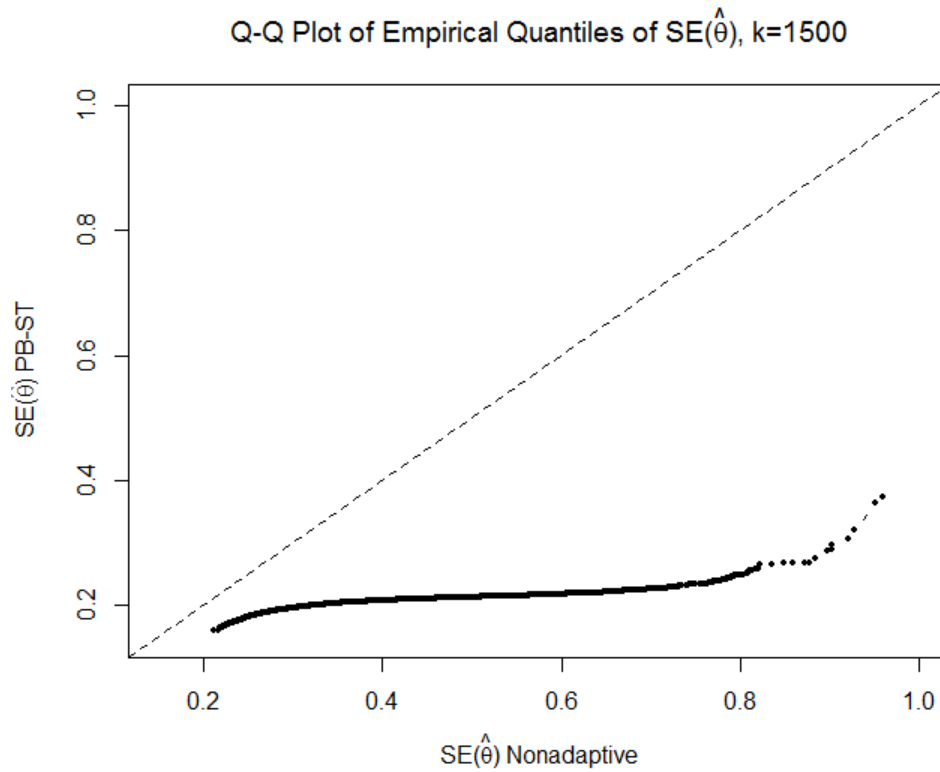


Figure 34. Q-Q plots of $SE(\theta)$ estimate distributions: Nonadaptive vs PB-ST ($k=1,500$)

Figures 35-37 are Q-Q plots comparing the performance of UE-ST and PB-ST. At its absolute best, UE-ST can indeed produce standard errors equally precise to those from PB-ST, if not more precise. However, the absence of explicit mechanisms for distributional control (as seen in PB-ST) or absence of concerns for overexposure (as seen in CAT) inevitably mean that UE-ST will have to administer suboptimal items in suboptimal circumstances. This finding is in accordance with the earlier analysis examining how $SE(\theta)$ and θ are related across methods, where the UE-ST item selection strategy has multiple potential vulnerabilities that differ based on examinee ability. As item bank size increases, PB-ST is capable of taking advantage of the larger depth of item selection, whereas the UE-ST uniform exposure mandate causes it to react similarly to Nonadaptive with respect to both item bank size increase as well as the degree of

incongruity between the population ability distribution and item bank properties. It cannot take advantage of added good items because it must also use the bad ones (and the subjects that do not need the good items will get them, while the subjects who do need them will get the bad items), so no significant improvement in performance is made as item bank size increases; and by virtue of having to administer all items equally frequently, it is inflexible and has no corrective measure for a circumstance where the item bank was designed for a different population or other content demands. Without the benefit of hindsight or comparable experience, one might find it curious that one method with completely non-adaptive selection and another with adaptive selection (albeit, with rigid exposure constraints) have very similar weak points, even if those weak points exist for different underlying reasons.

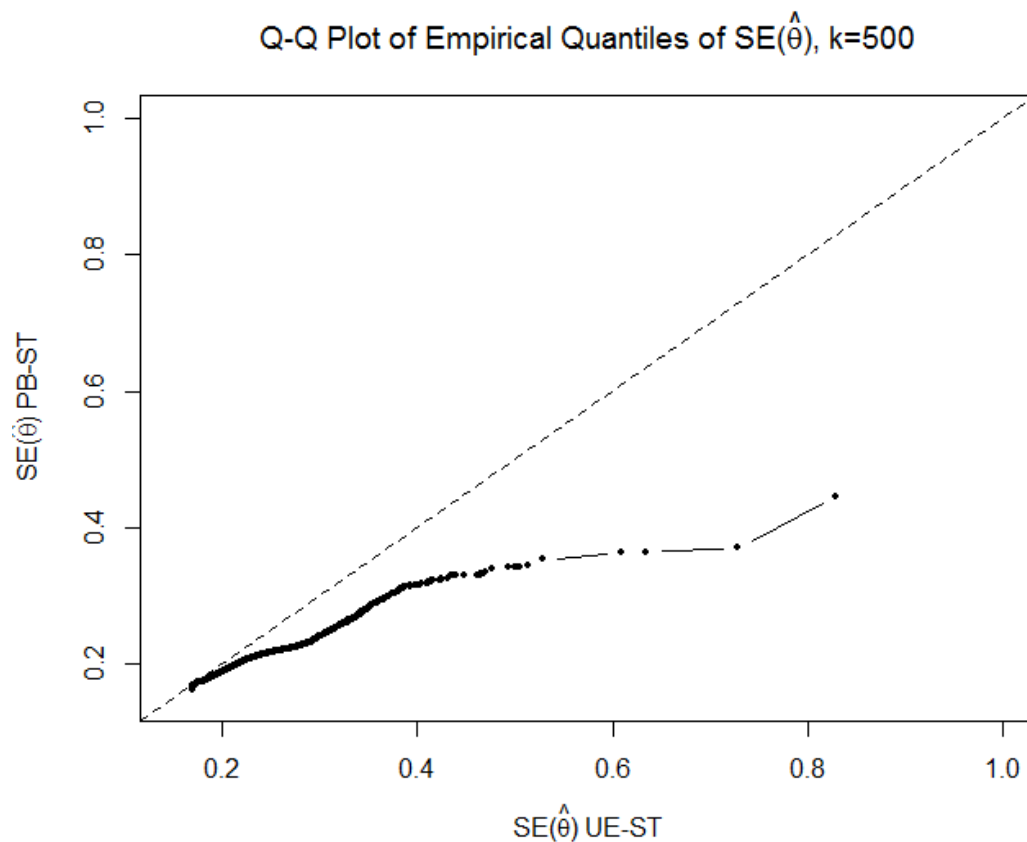


Figure 35. Q-Q plots of $SE(\theta)$ estimate distributions: UE-ST vs PB-ST ($k=500$)

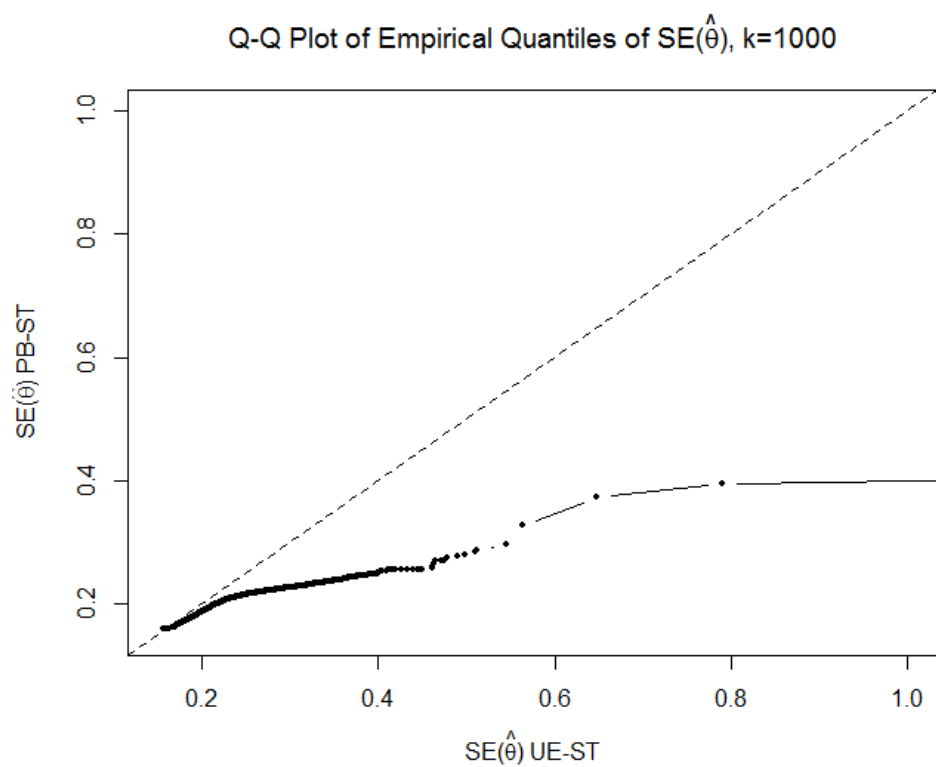


Figure 36. Q-Q plots of $SE(\theta)$ estimate distributions: UE-ST vs PB-ST ($k=1,000$)

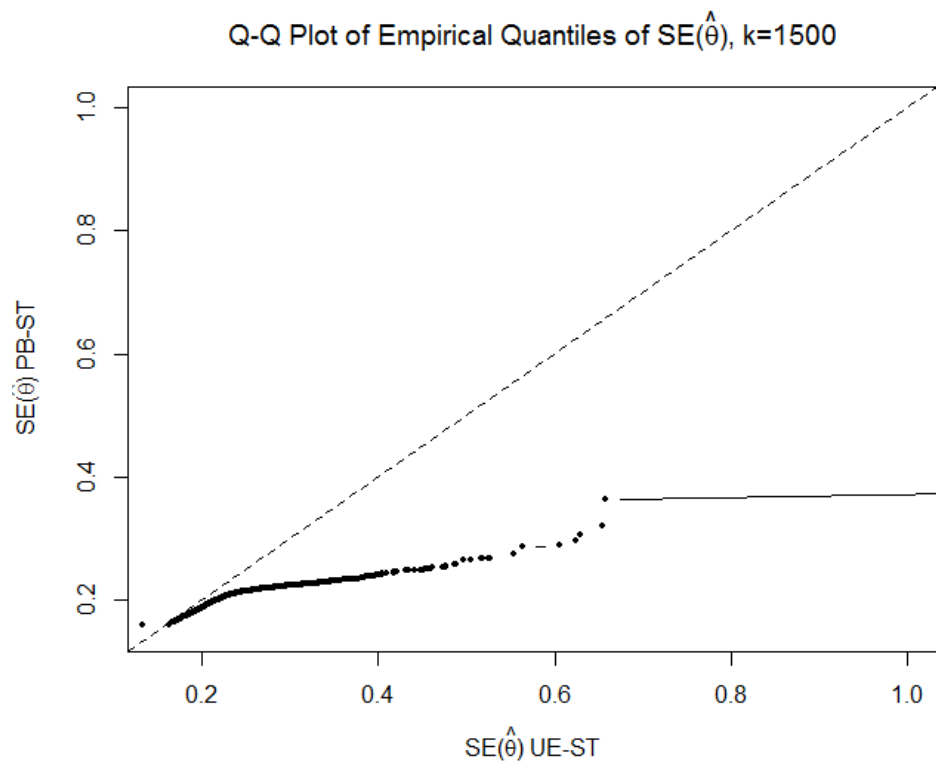


Figure 37. Q-Q plots of $SE(\theta)$ estimate distributions: UE-ST vs PB-ST ($k=1,500$)

Item Exposure

Section 1: Exposure in aggregate. Figures 38-41 present an interesting result; the CAT, PB-ST, UE-ST and Nonadaptive item selection frequencies were as expected, with the caveat that UE-ST was not perfectly uniform because it featured an integer program formulated with “greater than or equal to” rather than “greater than” in the maximum exposure part of problem formulation as a prophylactic measure for dealing with potential fringe cases in the estimation process. As a result, the effective maximum exposure was one count higher than it otherwise would be. A brief simulation examining the effect of this decision concluded that it was trivially, if at all, impactful as a deviation from normal procedure.

It appears to be the case that the PB-ST selection method did not use a certain percentage of the available items because the introduction of precision into the integer program no longer forced uniform item selection frequency. This admittedly is a comparatively much smaller percentage of the item bank that CAT typically discards, but is still not a trivial exclusion. Further examination of the properties of items not frequently selected under the PB-ST criteria reveals a common-sense answer; weakly discriminating items are not used (see figure 38) because they provide less information. In the context of item exposure control this is a failure to fully control overexposure, though outside of the small minority of unused items the selection frequencies do generally adhere to the guideline of uniform exposure. This result is arguably sufficient to explain the superior average performance on ability estimation between UE-ST and PB-ST; it is a simple matter of test information and the “need” criterion being calibrated in a such a way (as a simple ratio) that certain types of less informative items go unused.

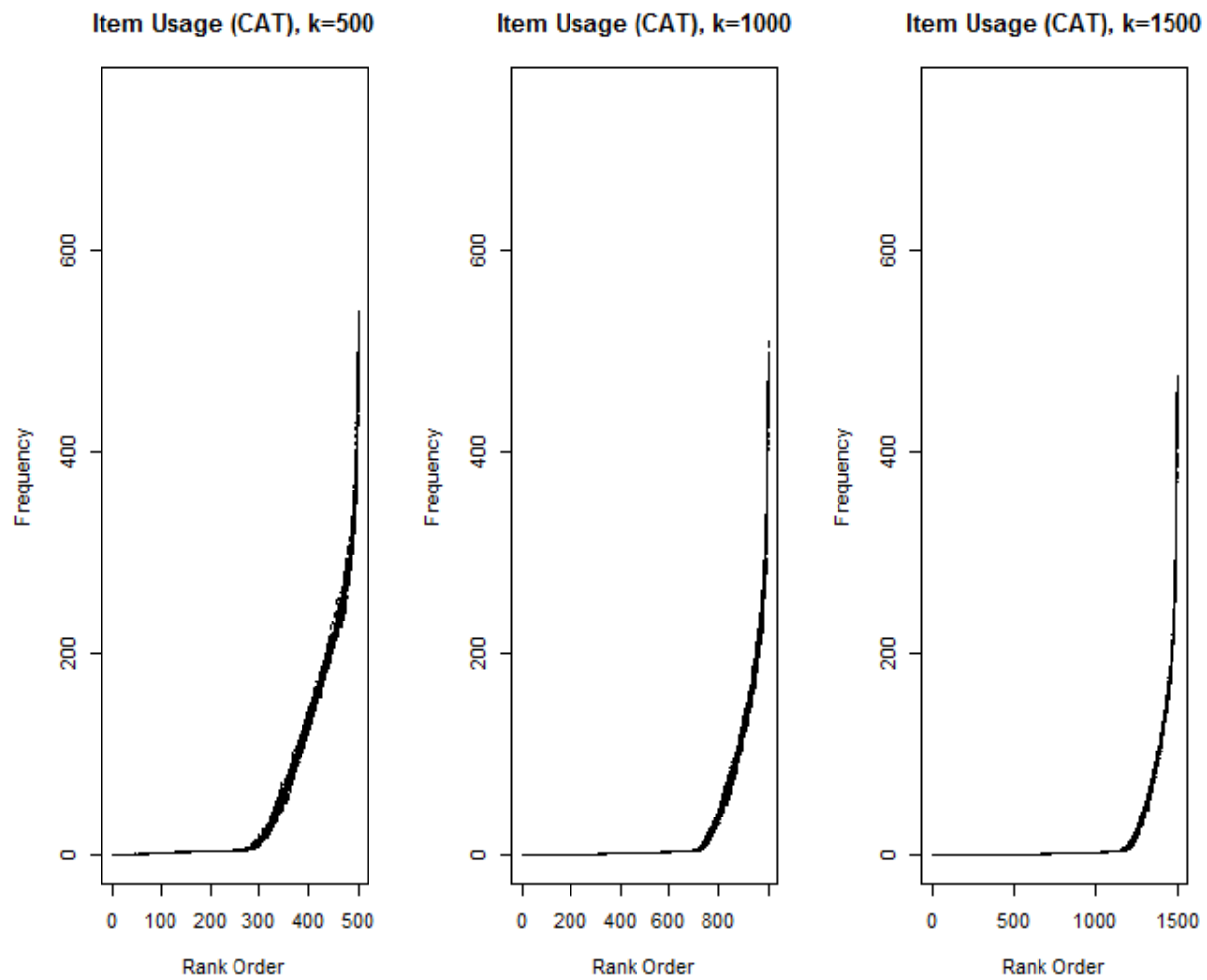


Figure 38. Rank order of frequency of item usage within CAT conditions

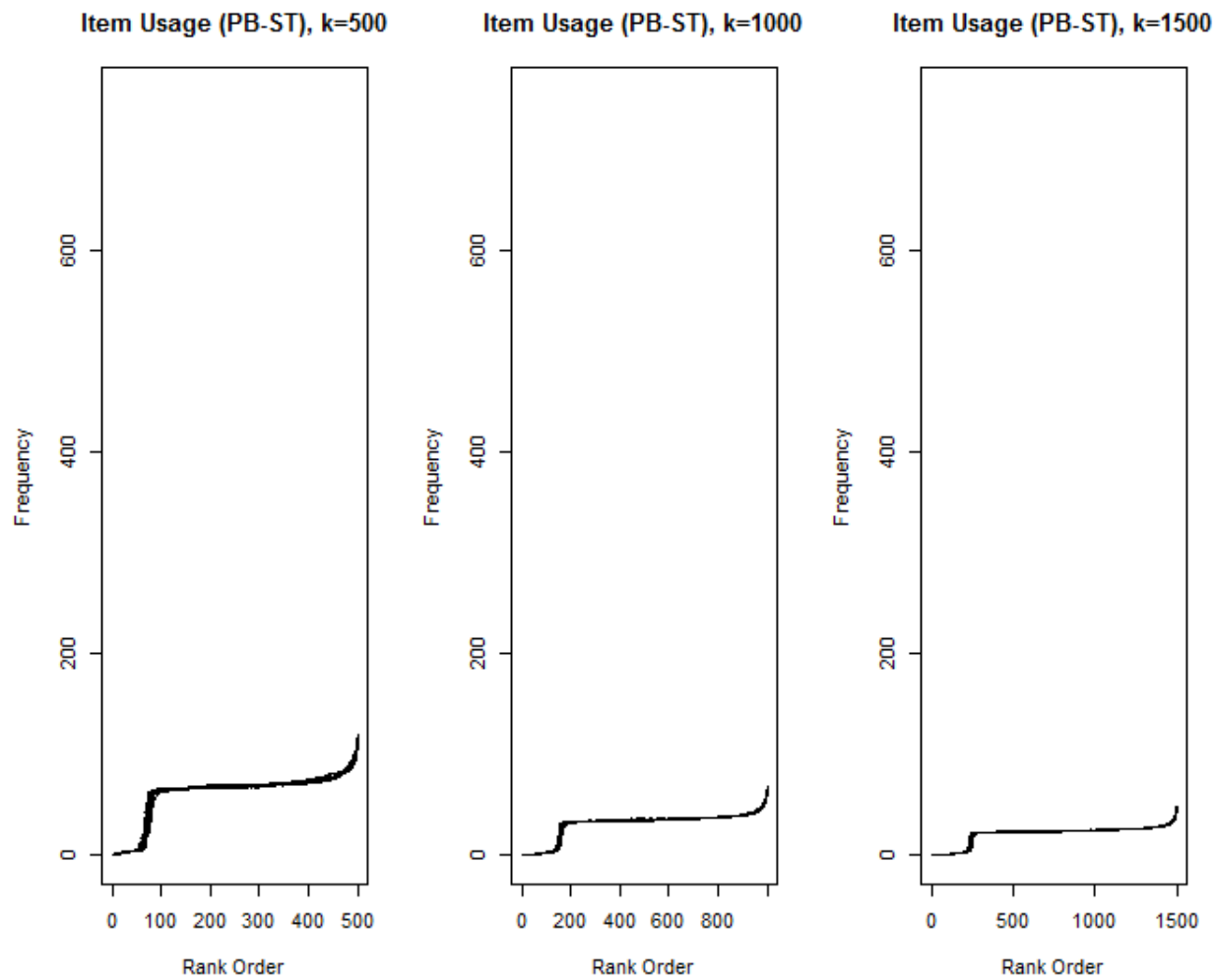


Figure 39. Rank order of frequency of item usage within PB-ST conditions

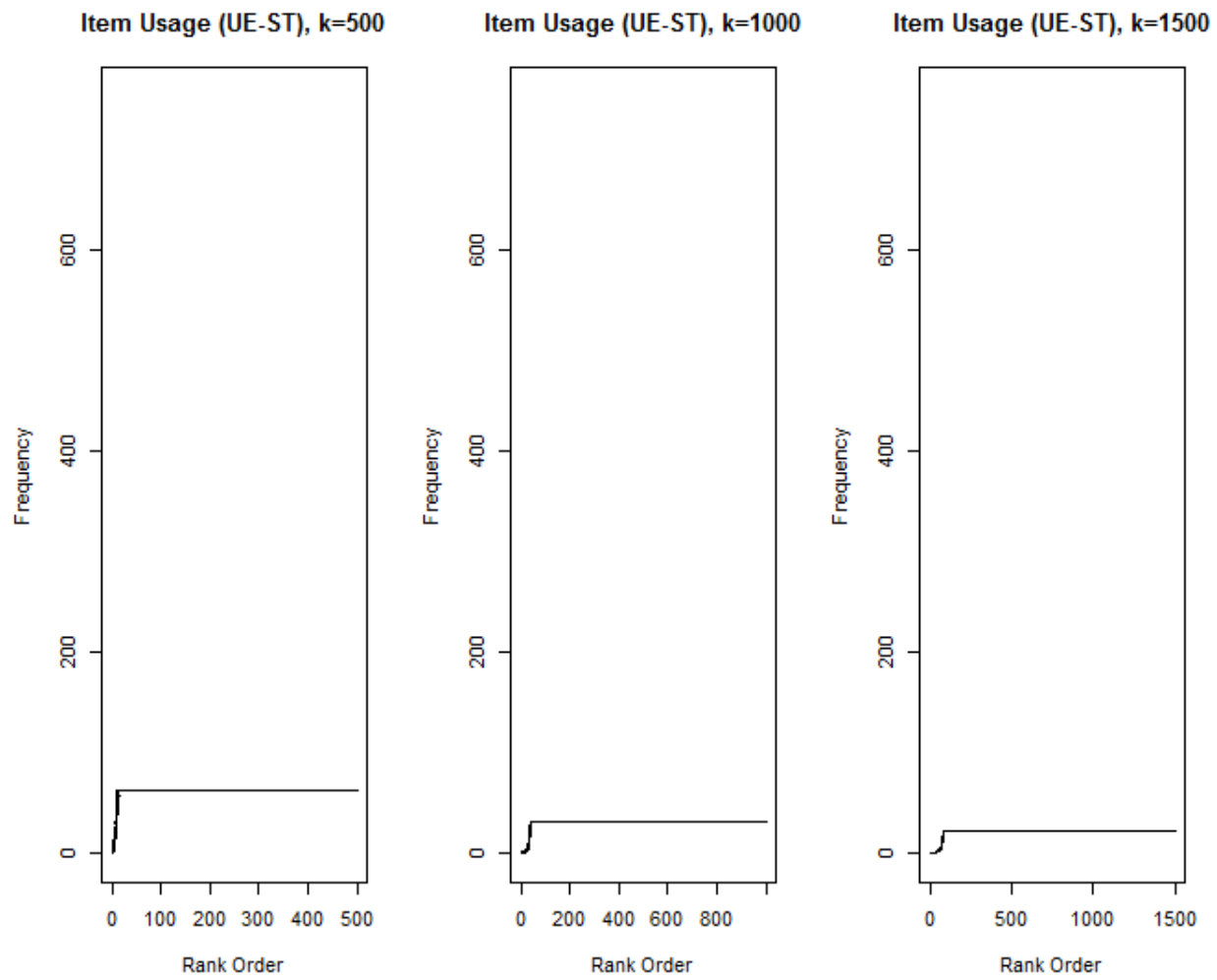


Figure 40. Rank order of frequency of item usage within UE-ST conditions

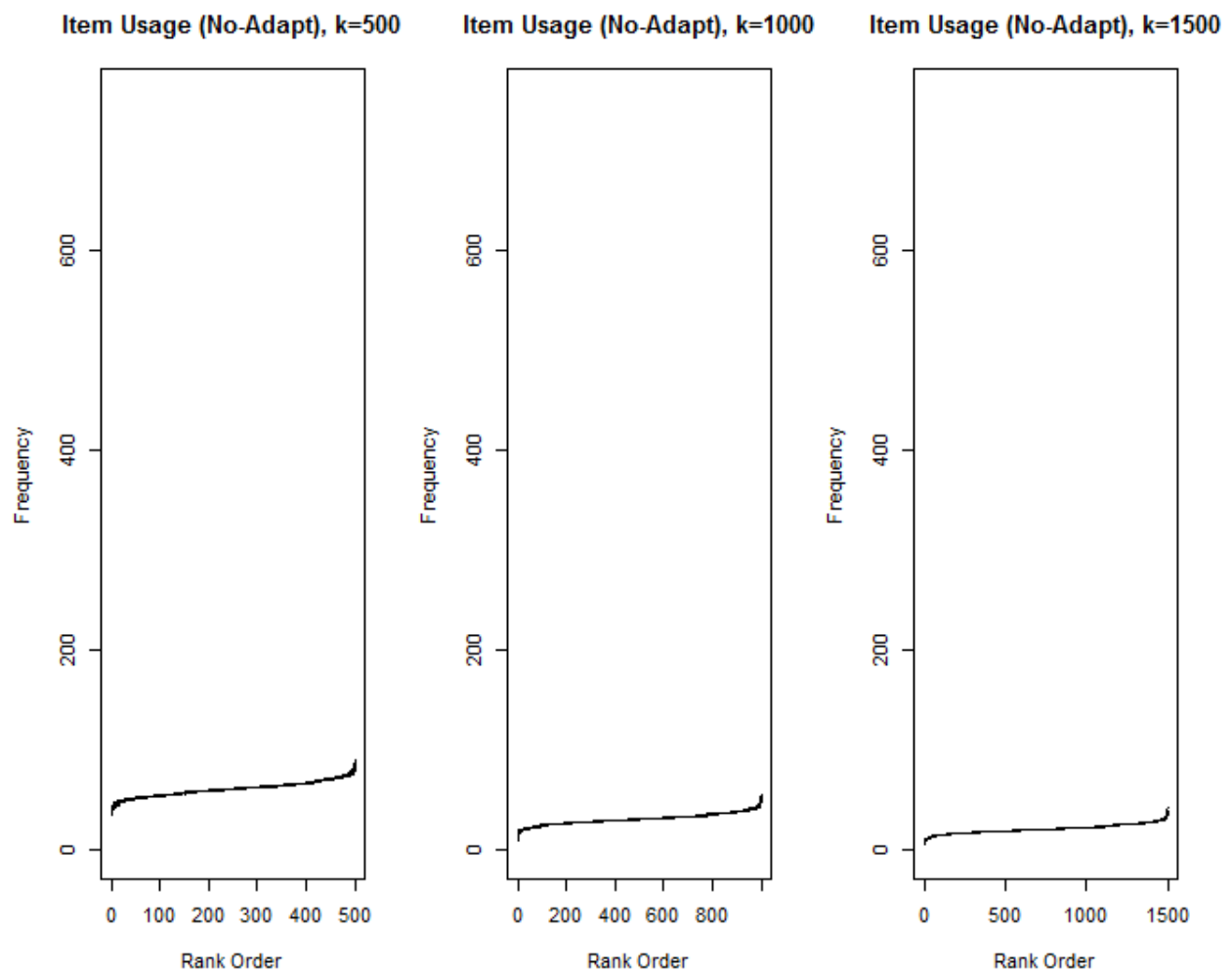


Figure 41. Rank order of frequency of item usage within Nonadaptive conditions

Additionally, these figures suggest that regardless of the item bank size, CAT only uses approximately 200 of the available items. This is because CAT's breadth of item selection is driven by the distribution of subject ability and test length, neither of which changes across the item bank size conditions.

Section 2: Exposure in relation to item properties. Displayed in figures 38-46 are a breakdown of how the individual parameters in the 3PL IRT model (the a , b , and c parameters) and item selection method influence probability of item selection.

Discrimination. There some trends that are readily apparent across all four figures (Figures 42-45); first, the discrimination parameter largely dictates item selection when it is permitted to do so (CAT/PB-ST) and does not when either uniformity of exposure is enforced (UE-ST) or item information is not performed adaptively (Nonadaptive). This predilection for selecting highly discriminating items is expected because the discrimination parameter exists as a multiplicative second-order term in the calculation of item information, resulting in higher values of information as it increases, and lower values when a is lower than 1. Second, across the item bank size conditions, there appear no differences in the general pattern of overexposure for any of the methods. As the item bank increases in size but sample size and test length remain the same, the expected exposure rate for uniform exposure decreases, which is why the patterns largely are proportionally similar (as demonstrated in the rescaled versions of the PB-ST figures).

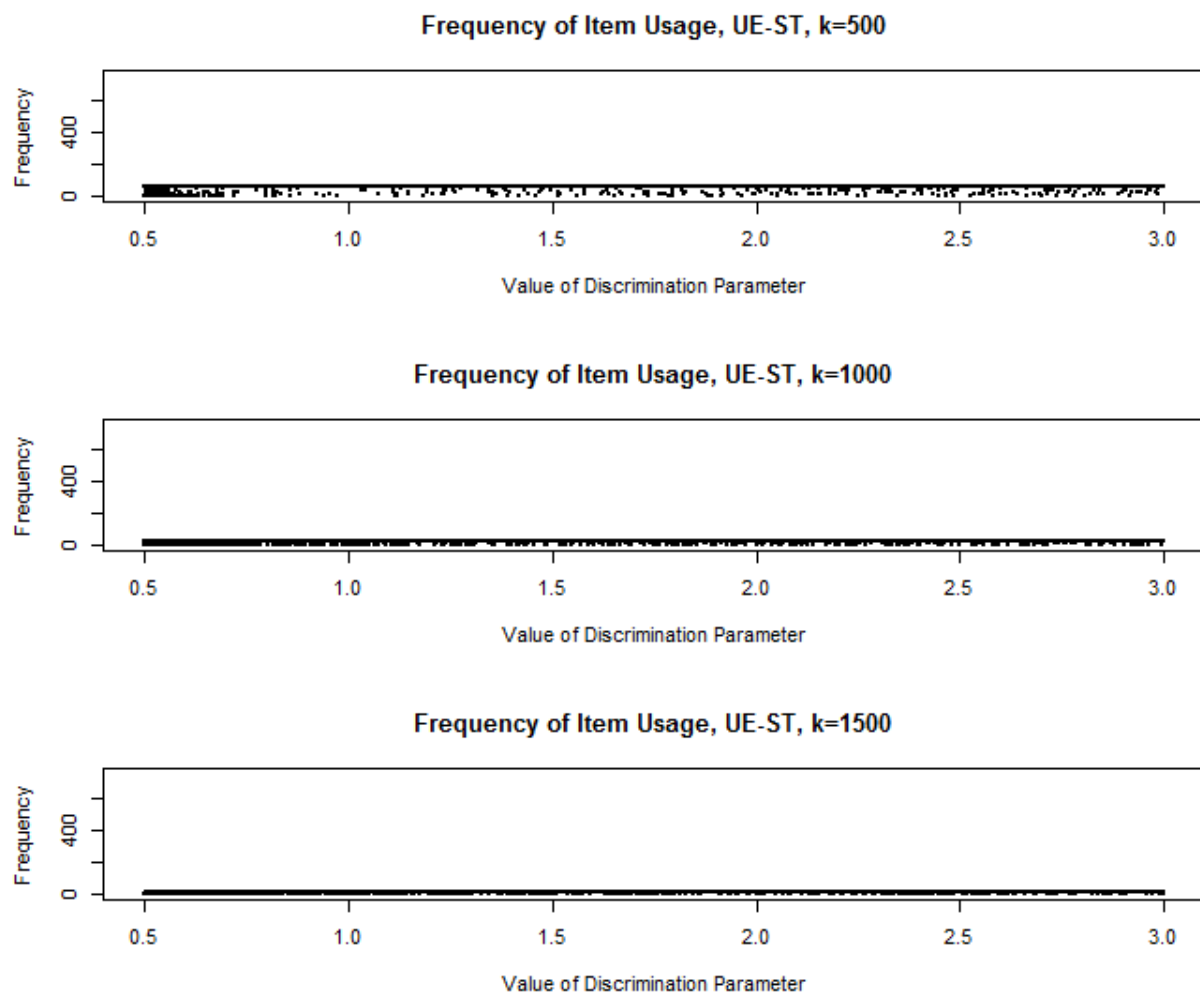


Figure 42. Discrimination parameter and frequency of item usage within UE-ST conditions

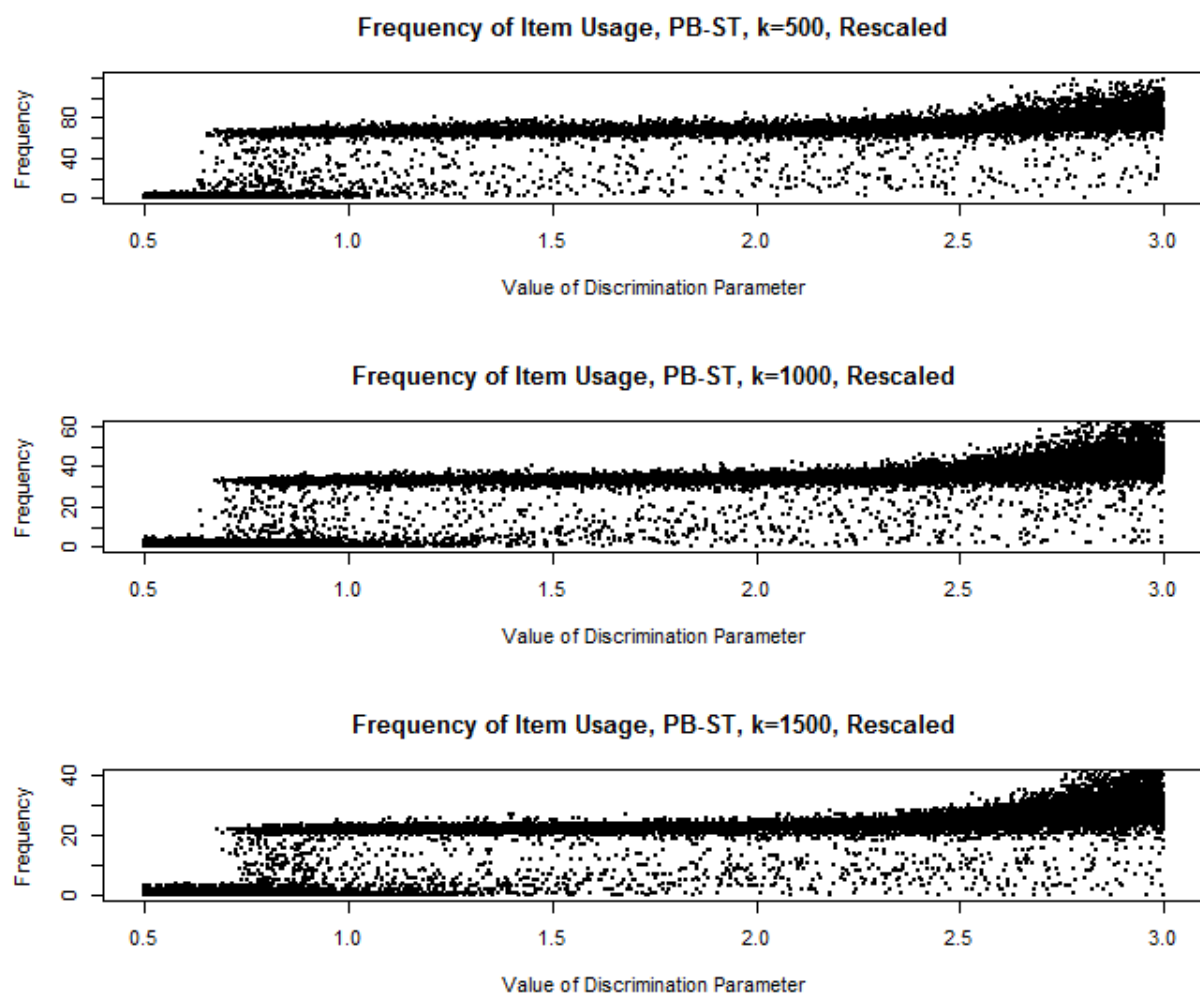


Figure 43. Discrimination parameter and frequency of item usage within PB-ST conditions

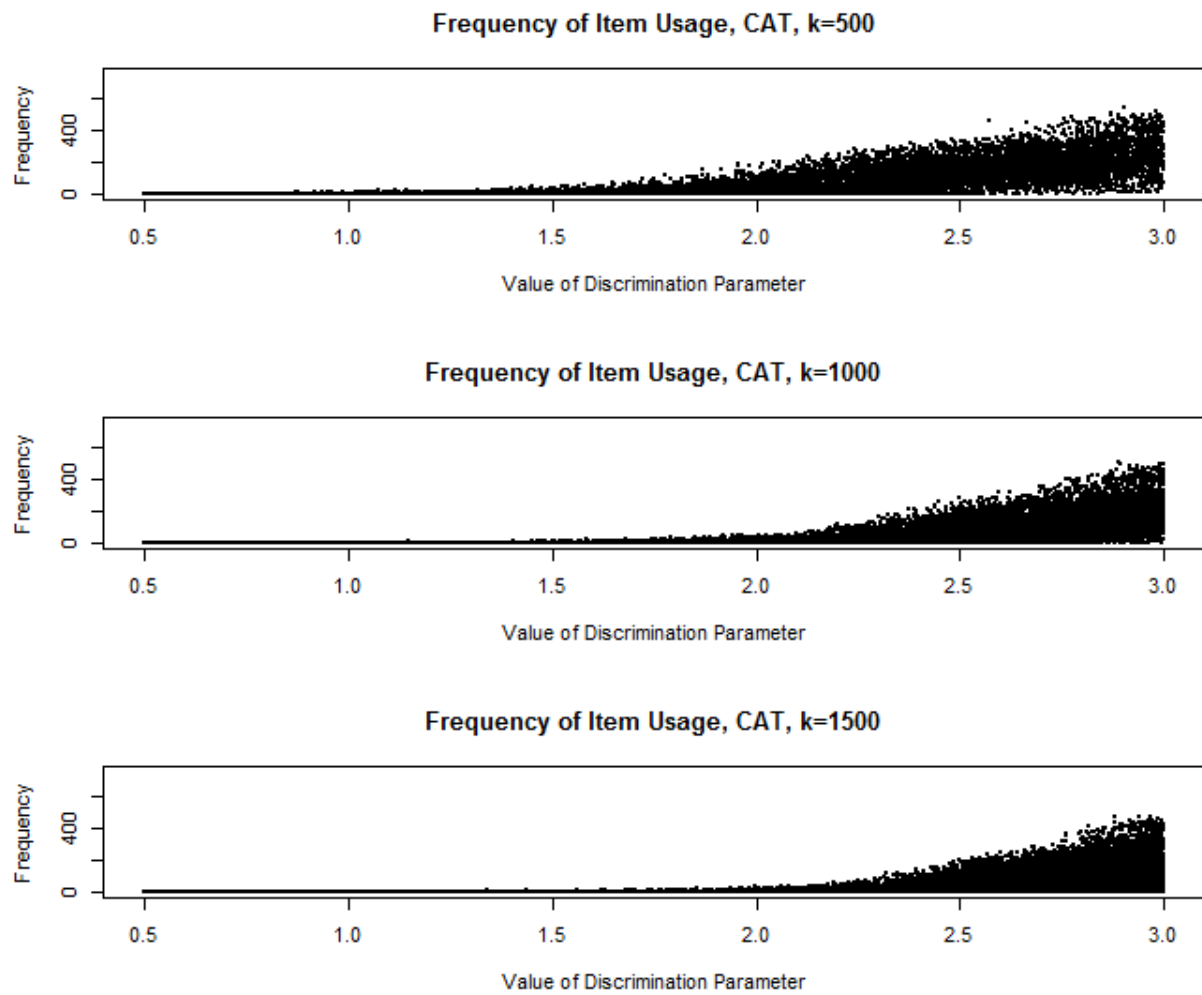


Figure 44. Discrimination parameter and frequency of item usage within CAT conditions

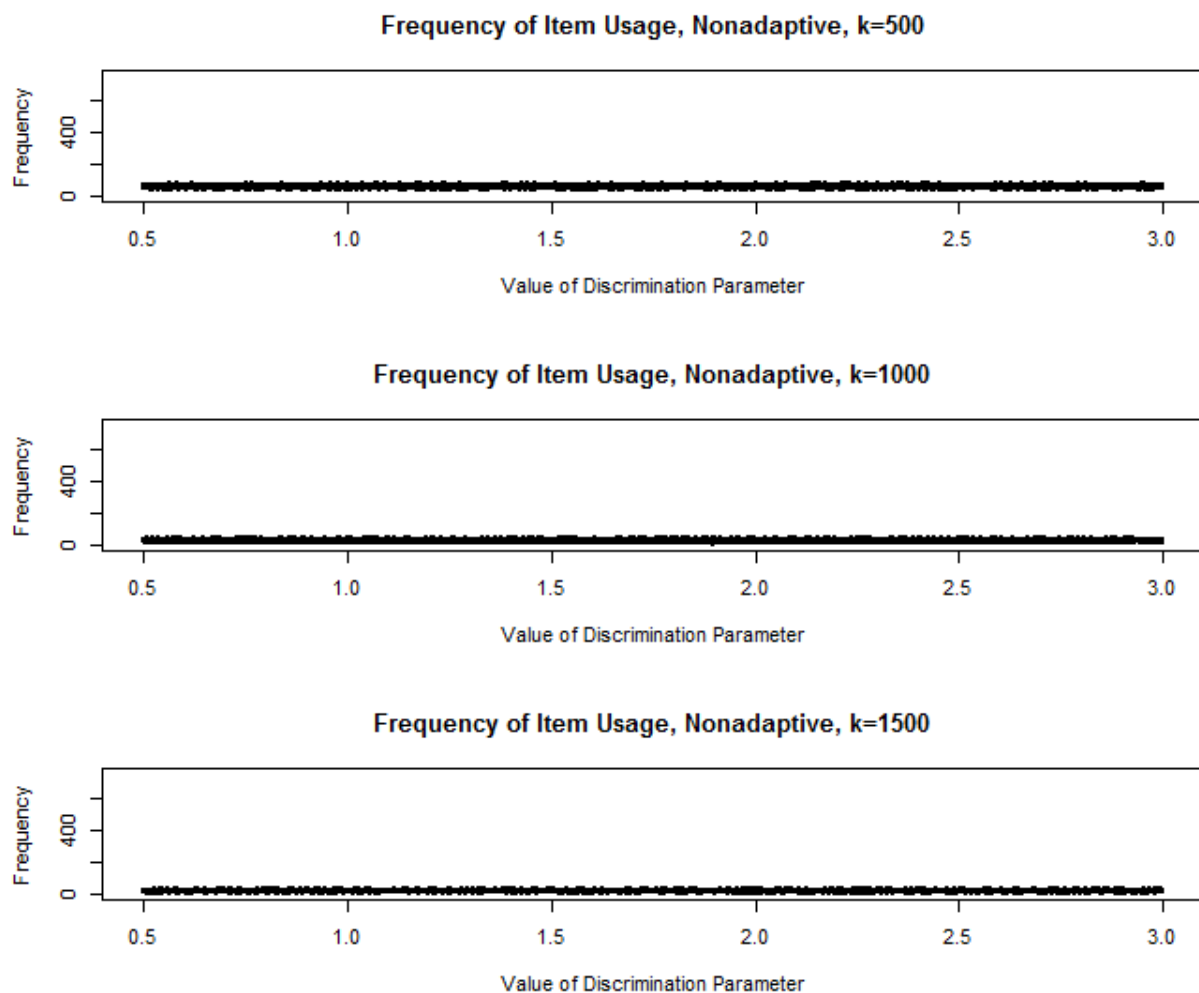


Figure 45. Discrimination parameter and frequency of item usage within Nonadaptive conditions

Difficulty. The value of the b parameter does not appear to substantially influence probability of item overexposure (Figures 46-49). In CAT, these figures essentially depict the distribution of b parameters in the item bank and ability in the population (both normally distributed, in this case), with some comparatively overexposed items present and spread proportionally evenly across the latent continuum as a result of having a high discrimination parameter. The Nonadaptive results also are intuitively straightforward in that items administered non-adaptively should not show an exposure pattern related to a psychometric property of the items. In the case of UE-ST/PB-ST, underexposed items appear to be primarily

located outside of the central area in the latent continuum, with high difficulty items having a slightly higher probability of being underexposed as a result of how item information is calculated in the 3PL model,

$$I_j(\hat{\theta}_n) = a_j^2 \cdot \frac{(P_j(\hat{\theta}_n) - c_j)^2}{(1 - c_j)^2} \cdot \frac{1 - P_j(\hat{\theta}_n)}{P_j(\hat{\theta}_n)}$$

rendering guessable low-difficulty items more broadly useful than guessable high-difficulty items. On a purely logical basis, one may deduce this property by noting that there is no reciprocated mechanism by which high-ability subjects can incorrectly answer low-difficulty items, in a manner attributable solely to randomness. The “inattention” parameter in the 4PL IRT model would be needed in such a case, and in the absence of this upper asymptote on the item response function to correspond with the existing lower one, the low-information, low ability, high-difficulty, high-guessing phenomenon that induces this asymmetry persists.

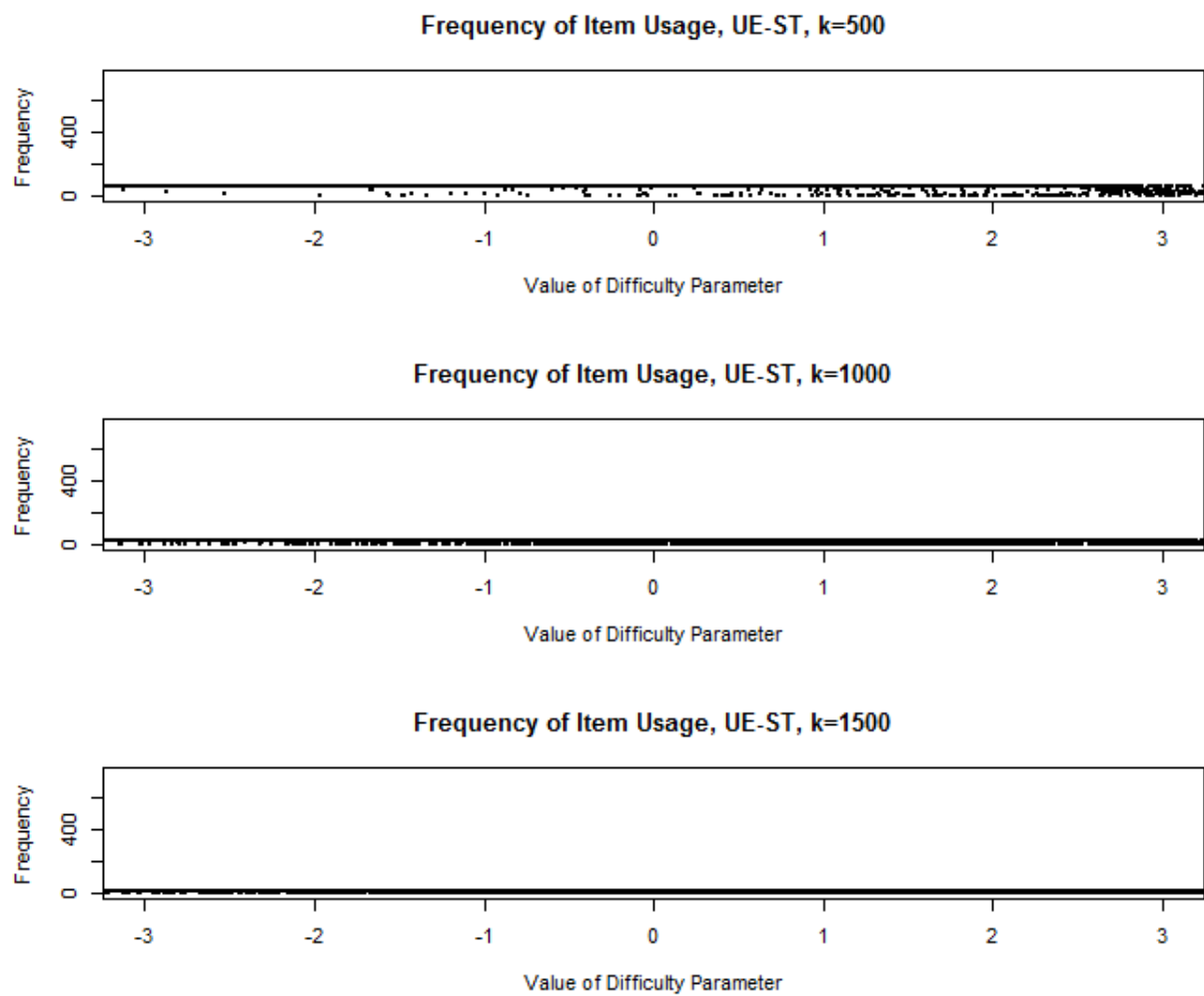


Figure 46. Difficulty parameter and frequency of item usage within UE-ST conditions

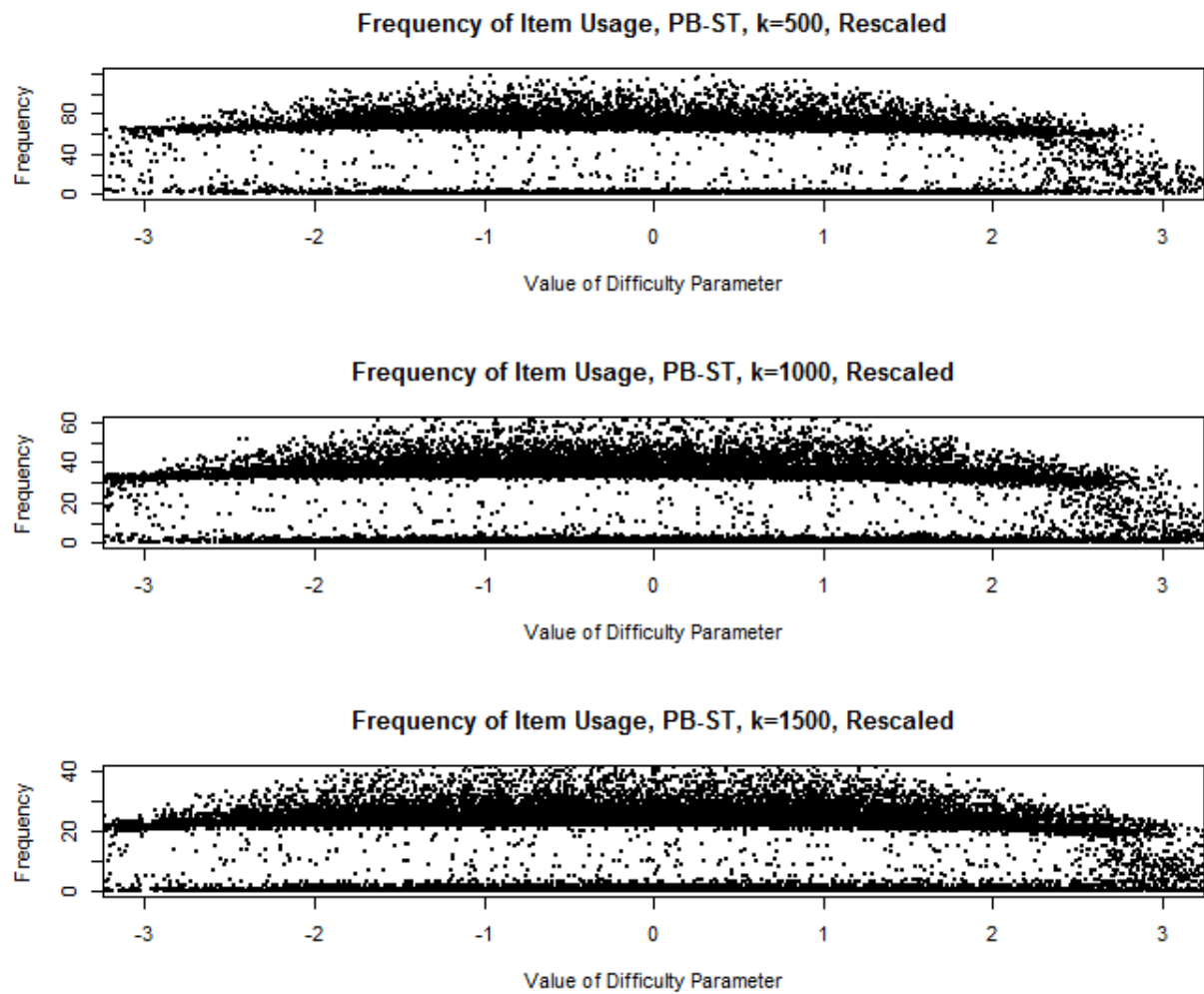


Figure 47. Difficulty parameter and frequency of item usage within PB-ST conditions

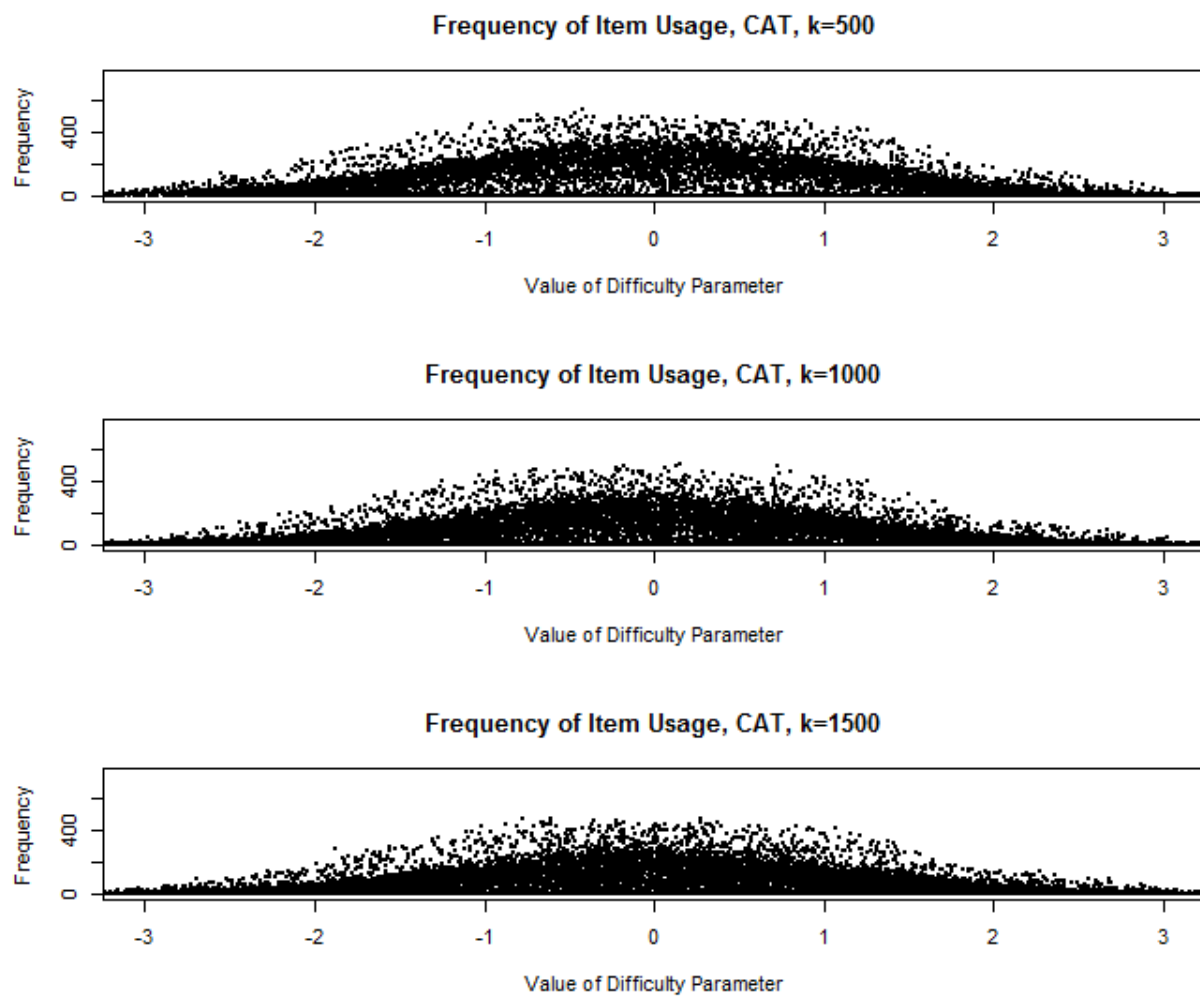


Figure 48. Difficulty parameter and frequency of item usage within CAT conditions

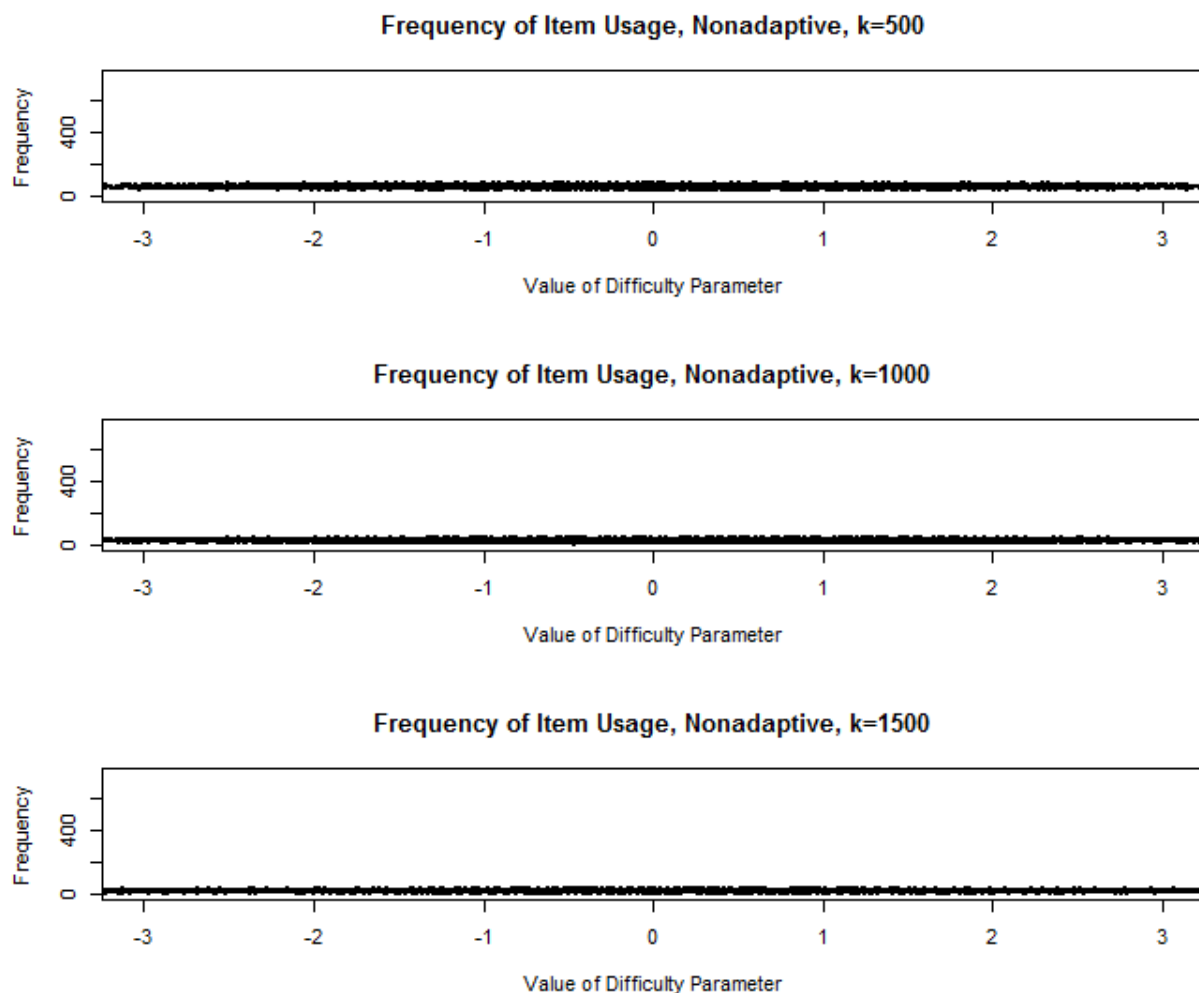


Figure 49. Difficulty parameter and frequency of item usage within Nonadaptive conditions

Guessing. As was the case with the discrimination parameter, it appears that c influences item exposure in the CAT or PB-ST methods and not the Nonadaptive or UE-ST methods. The incorporation of a lower asymptote into the item response function decreases the amount of information provided by the item as the asymptote is raised, so it is only natural that both the CAT and PB-ST methods favor items with smaller guessing parameters and underexpose items with higher guessing parameters. Unlike the dynamic for CAT and PB-ST with the discrimination parameters, items with higher guessing parameters are still frequently used, rather than going completely unused like low discrimination items. Refer to Figures 50-53.

This finding is more incidental than it is particularly indicative of the influential nature of 3PL item parameters on adaptive item selection; with guessing generated on \sim Uniform $[0,0.25]$ it is not nearly as impactful as it could be (modeling binary responses, as an example, would justify higher guessing parameters). Discrimination, on the other hand, features a sufficiently wide range of possibilities (generated on \sim Uniform $[0.5,3.0]$) that it can generate, from a standpoint based simply on item information, both “obligatory” to use and “pointless” to use items.

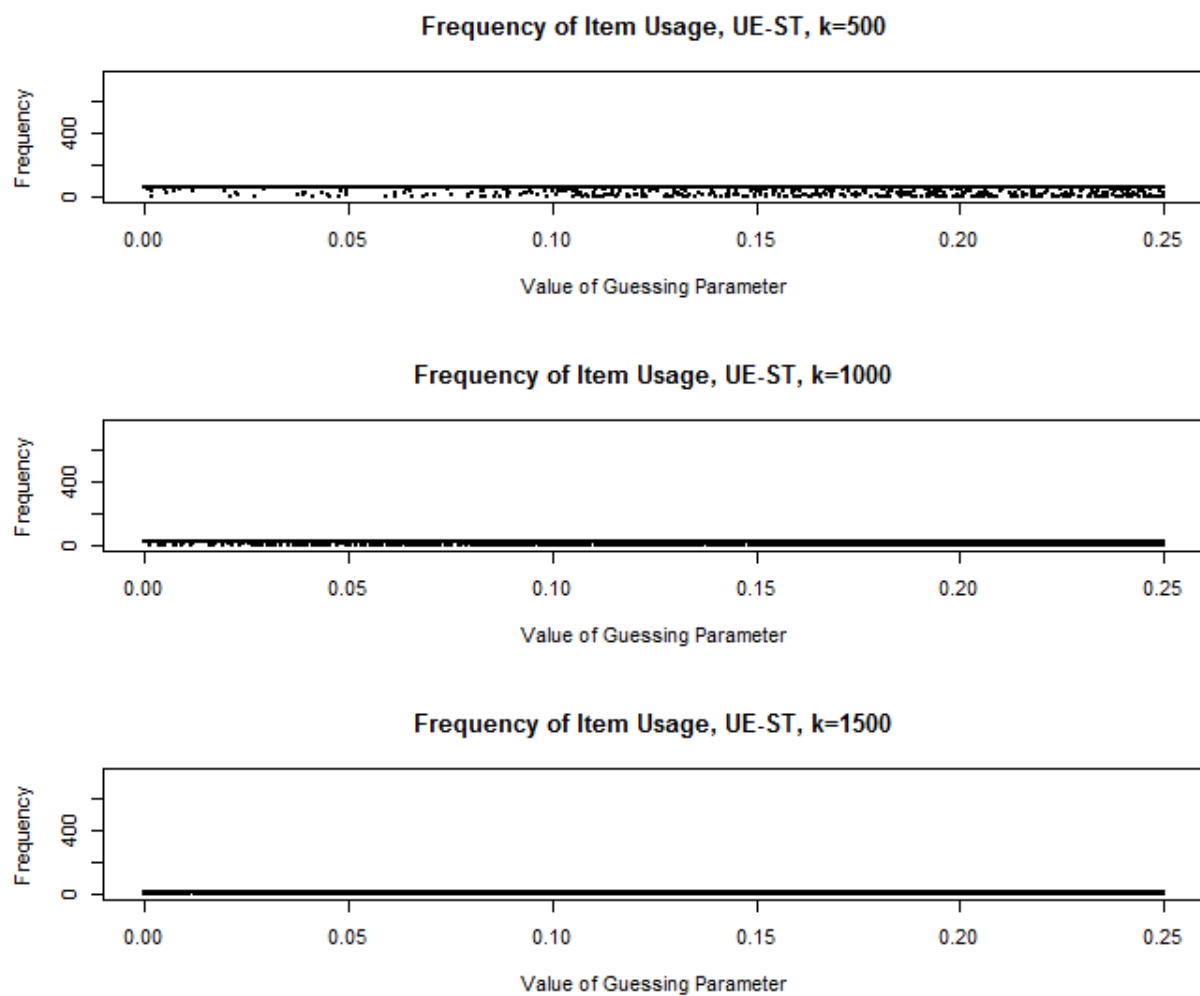


Figure 50. Guessing parameter and frequency of item usage within UE-ST conditions

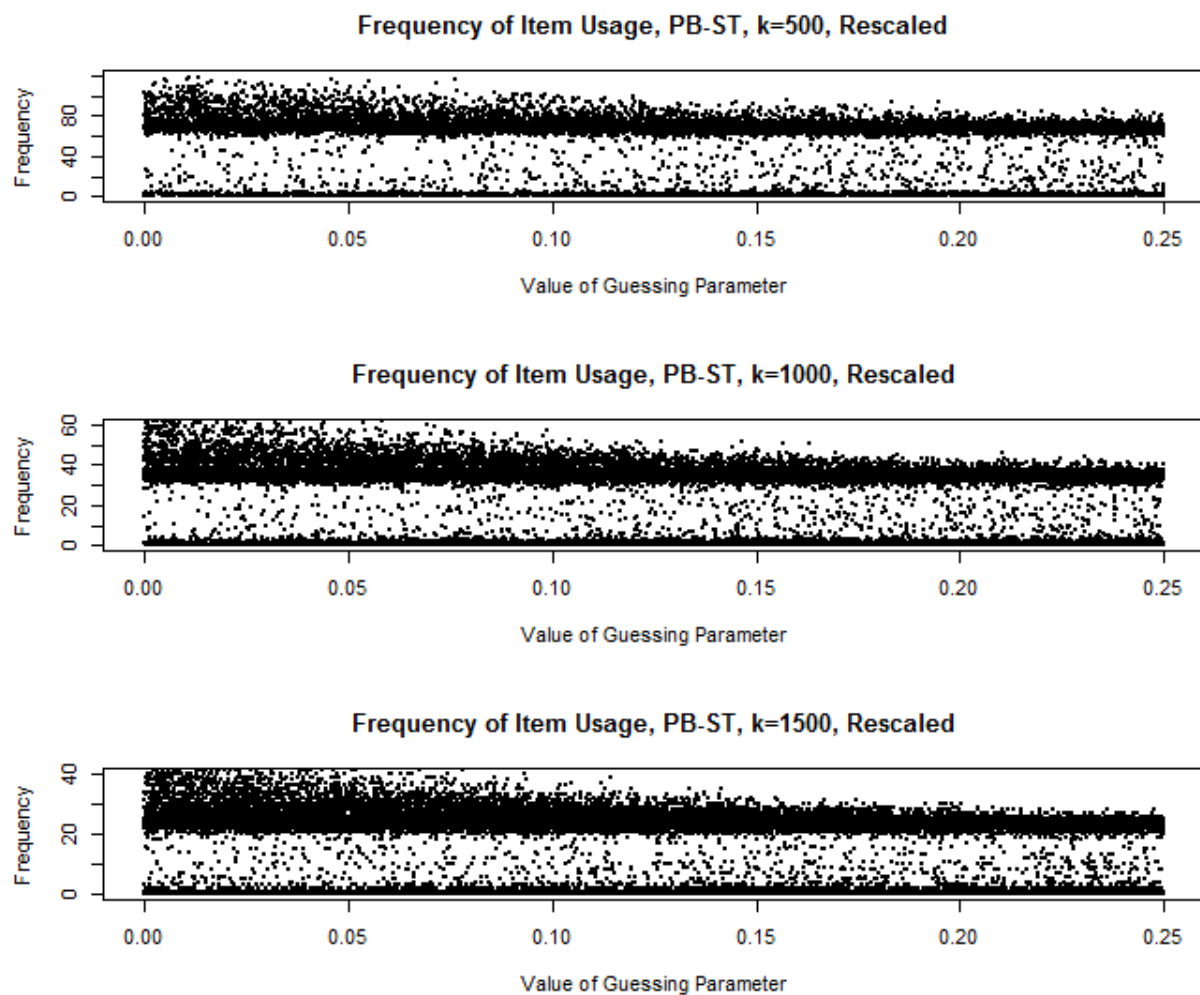


Figure 51. Guessing parameter and frequency of item usage within PB-ST conditions

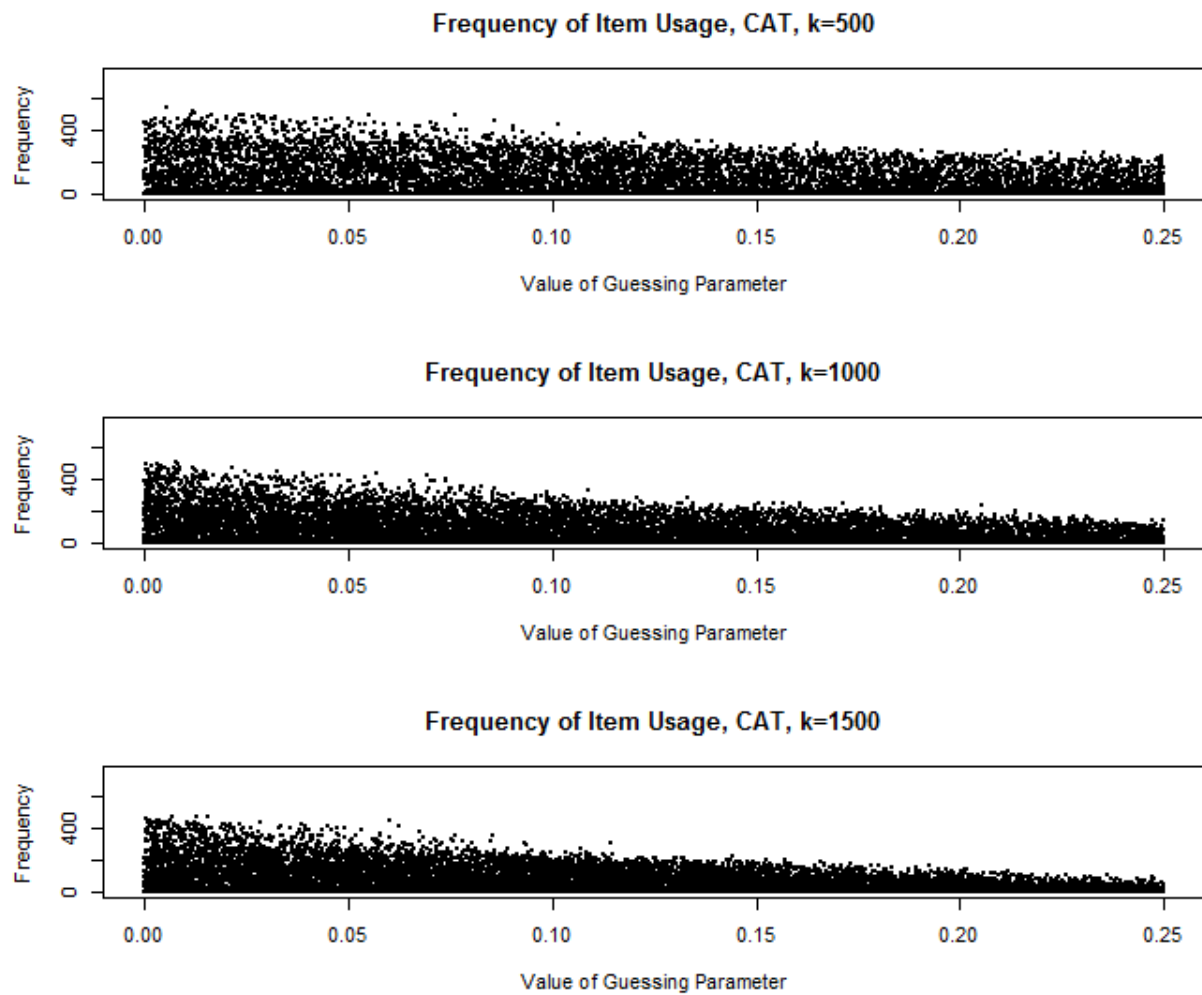


Figure 52. Guessing parameter and frequency of item usage within CAT conditions

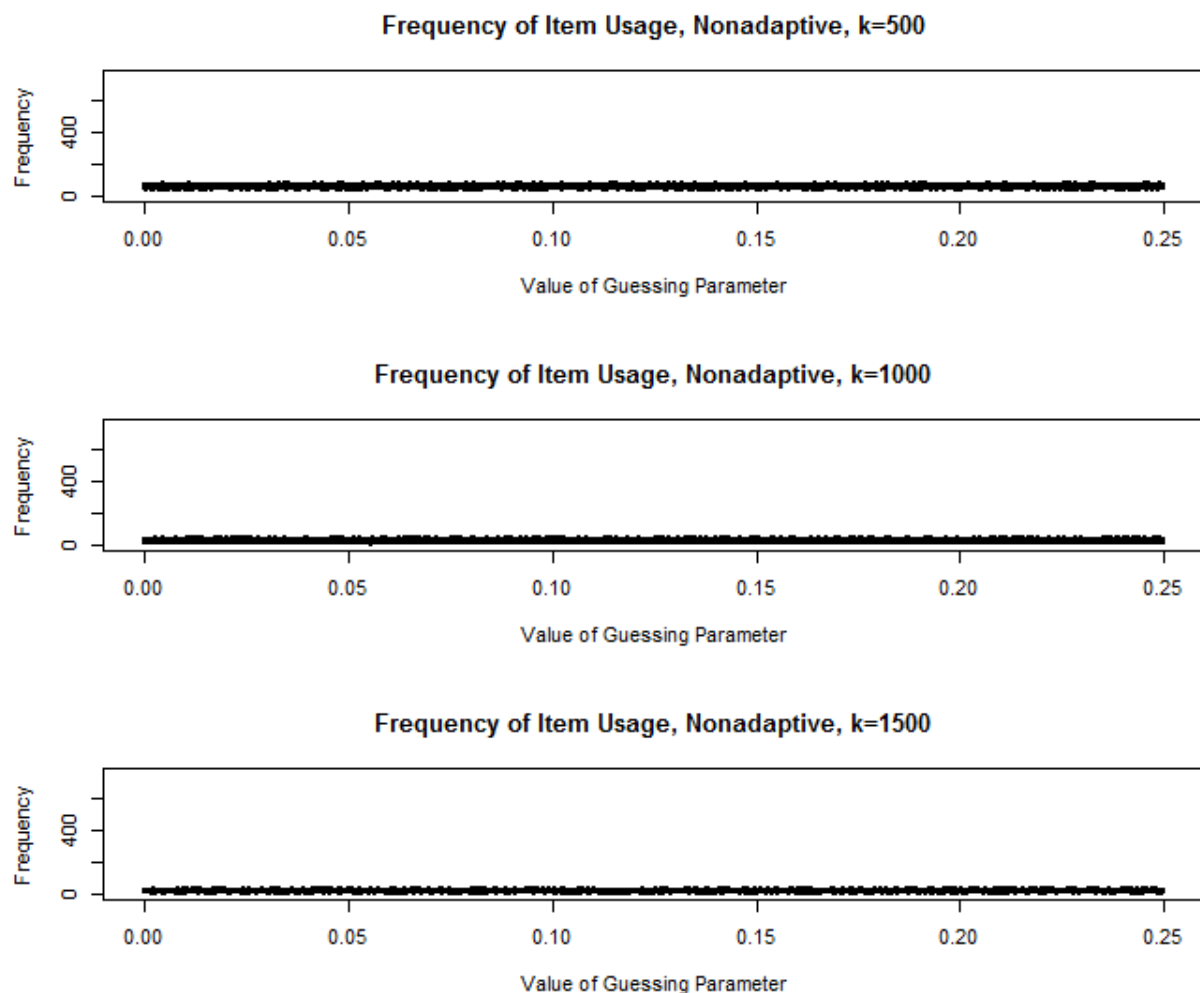


Figure 53. Guessing parameter and frequency of item usage within Nonadaptive conditions

Section 3: Vulnerability to item pilfering. Figures 54-56 are a set of scatterplots where percentage of shared items from two different methods is plotted by score on the latent trait. It was possible to calculate the shared items metric because each replication featured the same 750 subjects taking, in parallel, four separate tests. As such, the commonalities shared between these circumstances are (1) score on the latent trait and (2) the initial item selection, which was chosen randomly.

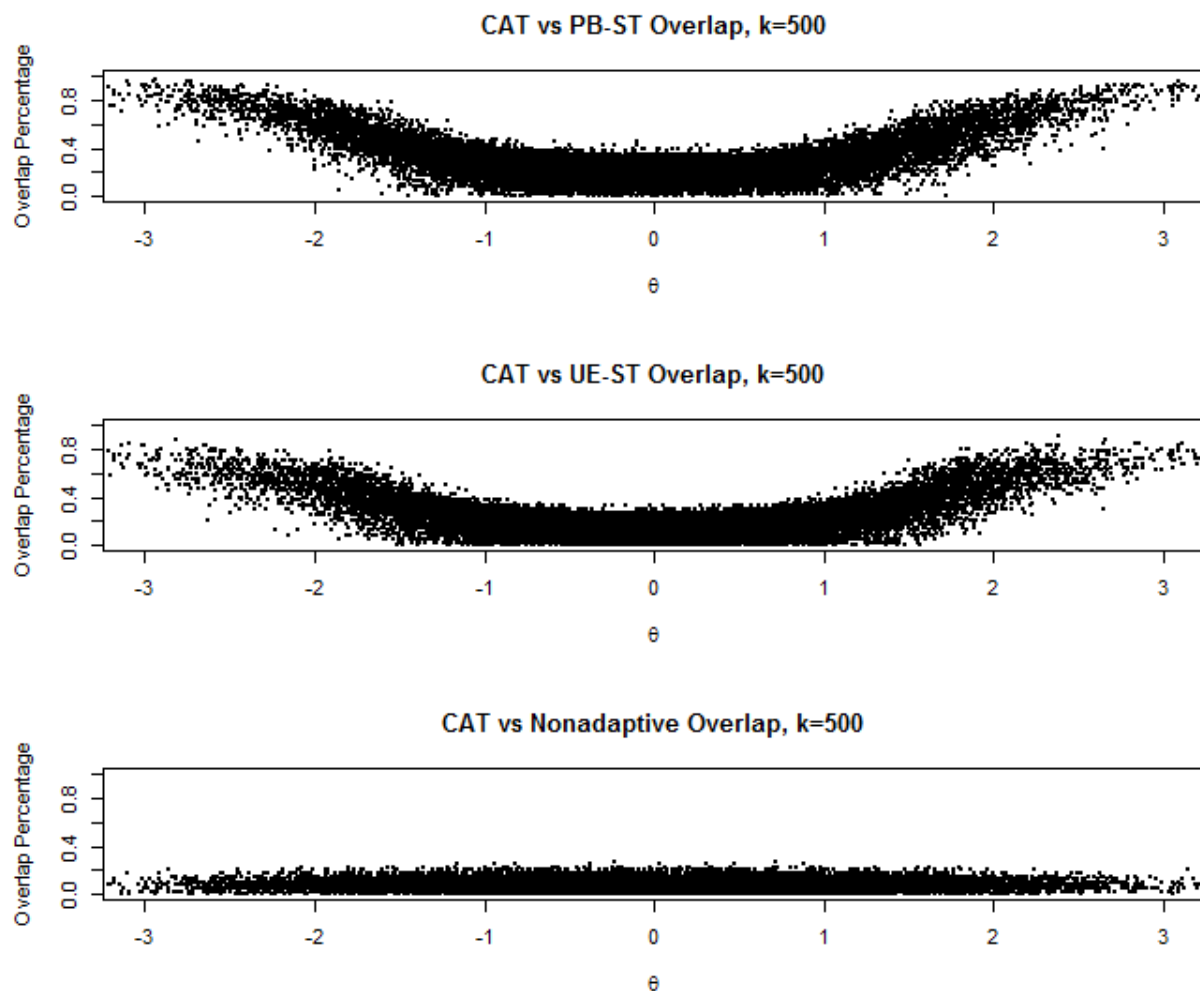


Figure 54. Pairwise comparison on shared items by score on the latent trait ($k=500$)

This means that any differences in subsequent item selections are the result of a few factors: (1) the item selection method, (2) “performance” on the item, as represented by a single random draw from a uniform distribution bounded at $[0,1]$, and (3) the exposure of items caused by other subjects (in UE-ST and PB-ST, where such values partially dictate the subset of available items). The second factor is rendered irrelevant on what one might call an asymptotic basis, in the sense that such differences in random number generation will not strongly influence a statistic obtained here from 26,250 simulated subjects. The third factor is essentially an

intended function for two of the methods in the first factor, and therefore we can state that any meaningful differences in item selection are attributable solely to the item selection method.

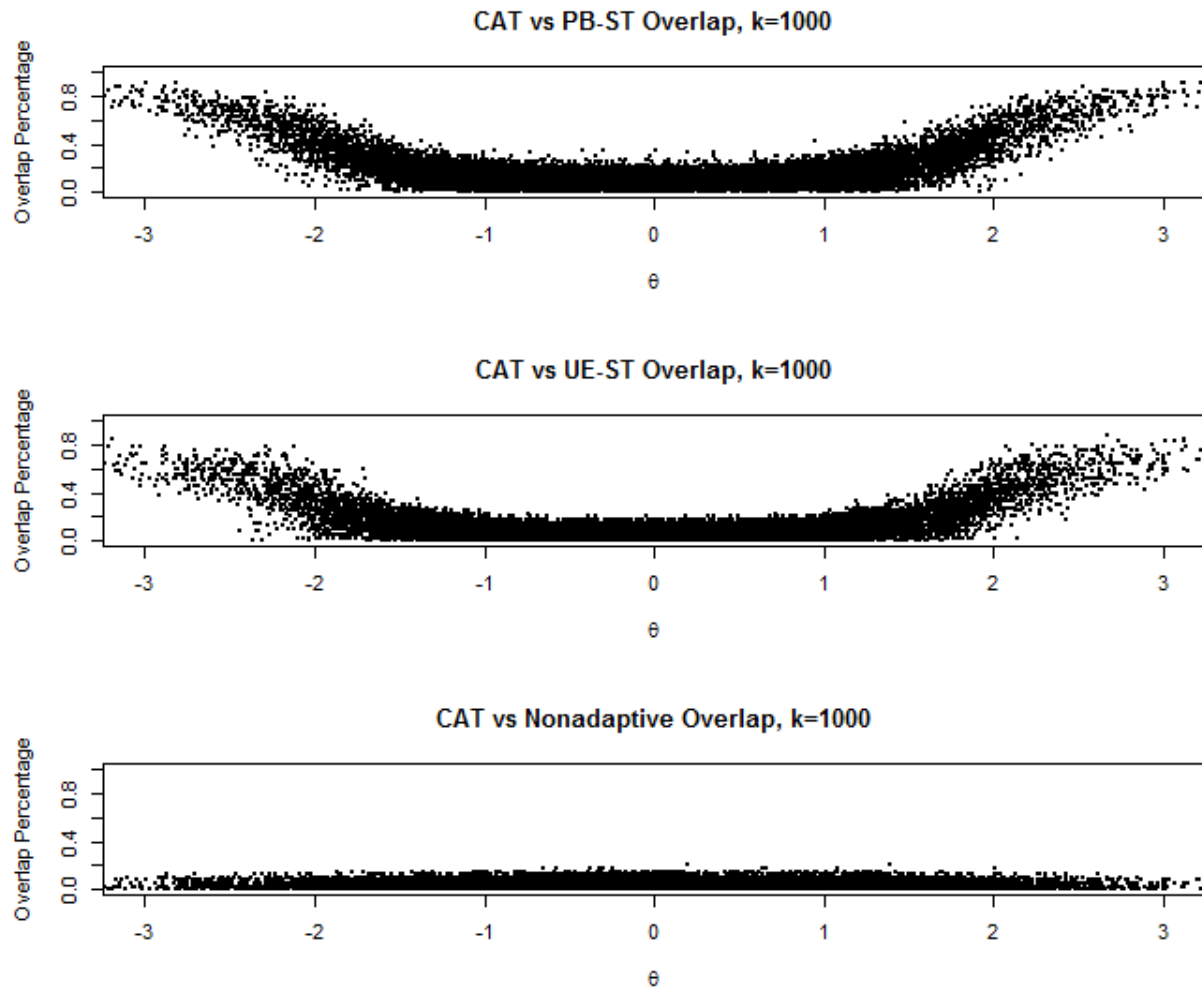


Figure 55. Pairwise comparison on shared items by score on the latent trait ($k=1,000$)

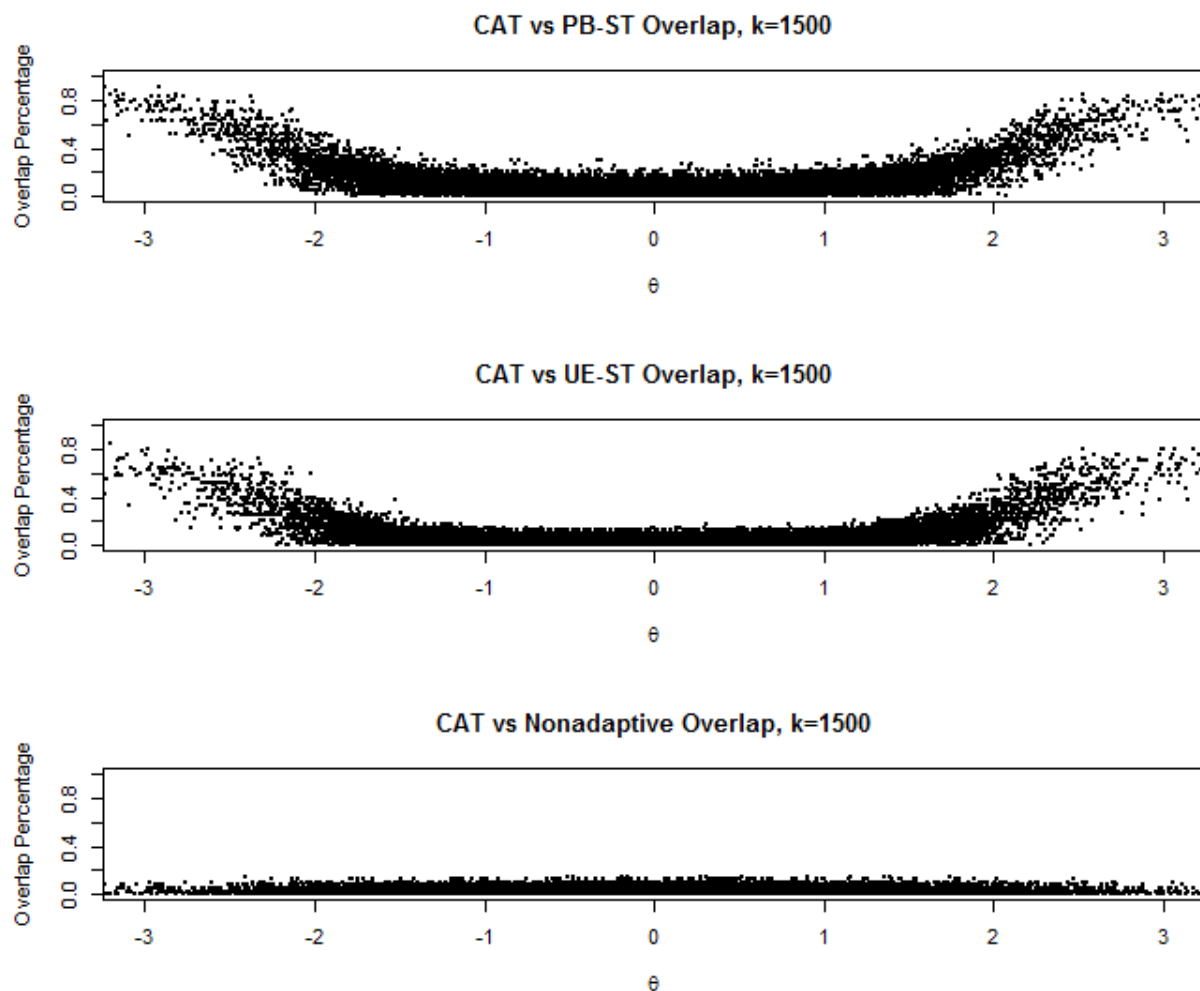


Figure 56. Pairwise comparison on shared items by score on the latent trait ($k=1,500$)

Figures 54-56 reveal somewhat expected information; the comparatively smaller number of viable items for subjects with abnormally low or high ability almost necessarily means that the probability of shared item selection between CAT and the UE-ST/PB-ST methods is higher. Although initially this may seem somewhat concerning that we observe higher shared percentages of items for precisely the subjects we do not want to share them (as dictated by the “item pilfering” component of earlier chapters), it is not a problem because the actual indication of whether a UE-ST or PB-ST successfully prevent item pilfering is comparing shared items with subjects of similar ability within the same test. In other words, it would be much more

concerning if there were high percentages of shared items for high-ability subjects within either the UE-ST or PB-ST methods. And this is what figures 57-59 address.

“High-ability” subjects were classified as being in the upper 10 percent of true scores on the ability measure (this percentile is somewhat arbitrary, but is a sufficiently reasonable value to be of use in our purpose here), and an overlap score was calculated as a mean percentage of overlap across all possible pairs of subjects in the high-ability group. More concisely, all pairs were laid out in an enumerative manner, overlap percentages were calculated, and then each unique subject obtained a single average overlap percentage. These average overlap percentages were used to form the densities by method in figures 57-59.

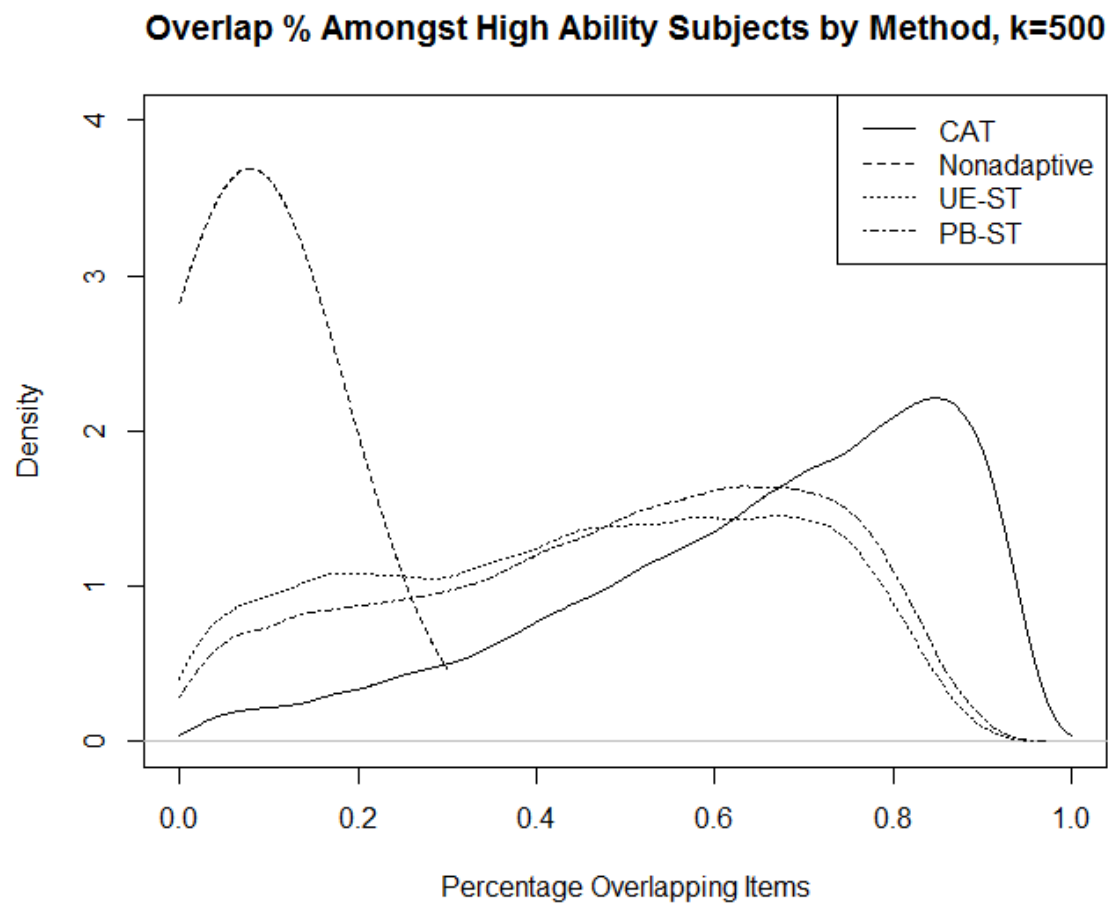


Figure 57. Overlap percentage amongst high ability subjects by method ($k=500$)

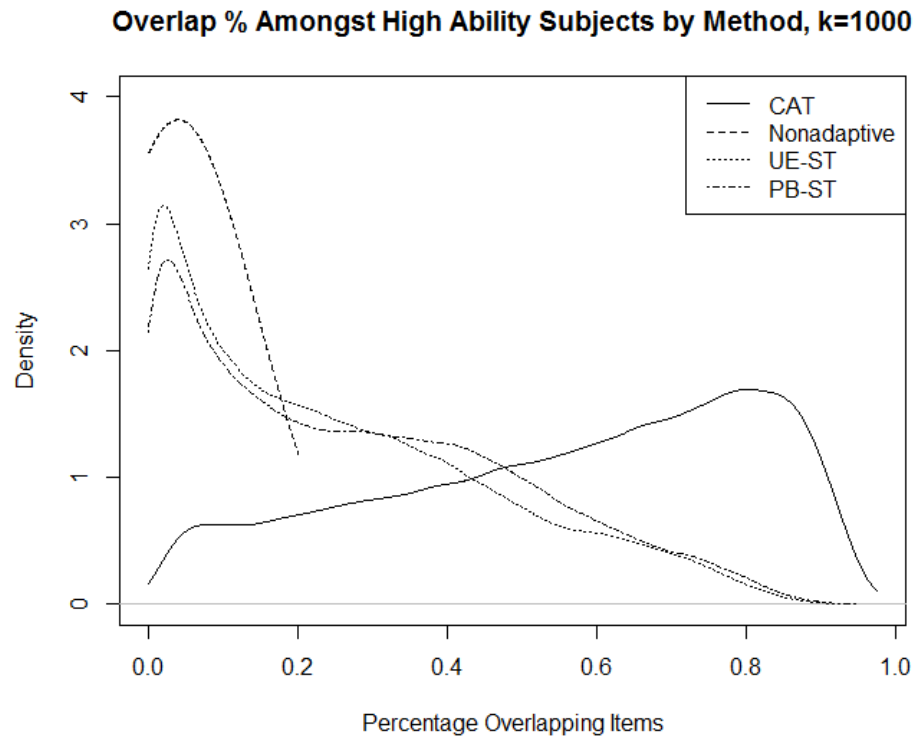


Figure 58. Overlap percentage amongst high ability subjects by method ($k=1,000$)

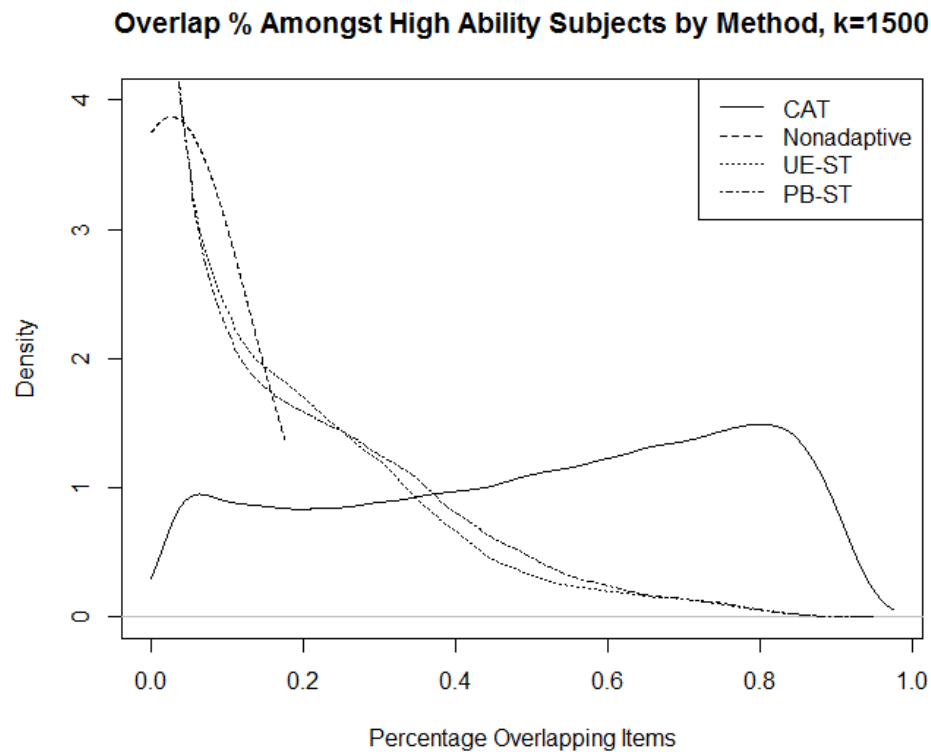


Figure 59. Overlap percentage amongst high ability subjects by method ($k=1,500$)

The result for the Nonadaptive method is not surprising, as any increase in value of overlap would necessarily be a result of increasingly unlikely random selection, so the small values are expected. It is difficult to confidently state any consistent differences between the densities represented by UE-ST and PB-ST, though it appears to be the case that PB-ST has slightly higher shared item percentages on average across the conditions. CAT unsurprisingly has the highest values and most consistent replication of shared items -- the method is implicitly designed to gravitate to a smaller number of higher quality items.

Perhaps the most interesting feature of figures 57-59 is the quite visually apparent change in distribution shape for the UE-ST and PB-ST values. The low item bank size condition, where only 500 items were available, in concert with other study parameters likely represents a scenario where the small number of available difficult items almost necessitates large overlaps in item selection for high ability subjects. Somewhere between item bank sizes of 500 and 1,000, however, there appears to be a point there are enough high difficulty items that it is possible to construct viable non-overlapping sets of high difficulty items for different subjects. This becomes evident in the item bank size 1,000 and 1,500 conditions, where the percentage overlap for high ability subjects in UE-ST and PB-ST appears to be slowly converging on the overlap rates seen in Nonadaptive, which is an ideal case when considering that uniform exposure and limited overlap in item selection (even across content and ability domains) is the intended prophylactic strategy for counteracting prospective item thieves.

Chapter V: Discussion

Part 1: General Observations

While there is some inherent value in observing, clearly and across many different aspects of the ability estimation and item exposure control processes, the effective performance of PB-ST relative to other methods, the most valuable piece of knowledge or information to come out of this simulation is the affirmation of the general viability of the study concept.

The concept is very simple; if you are going to implement a method that seeks to control item exposure, this necessarily means that at some point you will need to administer suboptimal items (in the context where optimal is defined as maximally informative). Every exposure control method does this; the mechanism by which suboptimal items are selected varies greatly, but the act itself is a unifying characteristic.

What has been done here is taking the necessary act, the administration of suboptimal items, and simply approaching it as a resource allocation problem where one mitigates the “damage” the act causes while maximizing the potential gain. We say that damage is minimized when a subject has a sufficiently precise estimate of ability that the subject does not need an informative item; the gain is maximized by, in concert with this tactical administration of suboptimal items, reserving items that are informative for circumstances where it is necessary to administer them.

The strategy effectively handles the problem of item pilfering by virtue of negating the unwinnable economic dynamic as revealed by Zipf’s law; as the entity administering the exam attempts to increase test security by increasing the size of the item bank, its costs increase exponentially while entities seeking to undermine the validity of the exam have their costs increase linearly. The results of the simulation indicate that this is no longer the case; the rank

order of item selection frequency and logarithmically transformed frequency are not linearly related in PB-ST, and item overlap for high-ability subjects is sufficiently limited that the item pilfering strategy is no longer efficient or, more fundamentally, it is questionably effective. As a partial homage to the graphs from earlier, and a partial demonstration that the item exposure dynamic is fundamentally changed with the implementation of precision-based shadow testing, see figures 60-62 wherein item exposure plots attempting to replicate Zipf's law (as seen prominently in Wainer 2000) are presented for the methods and conditions in this simulation.

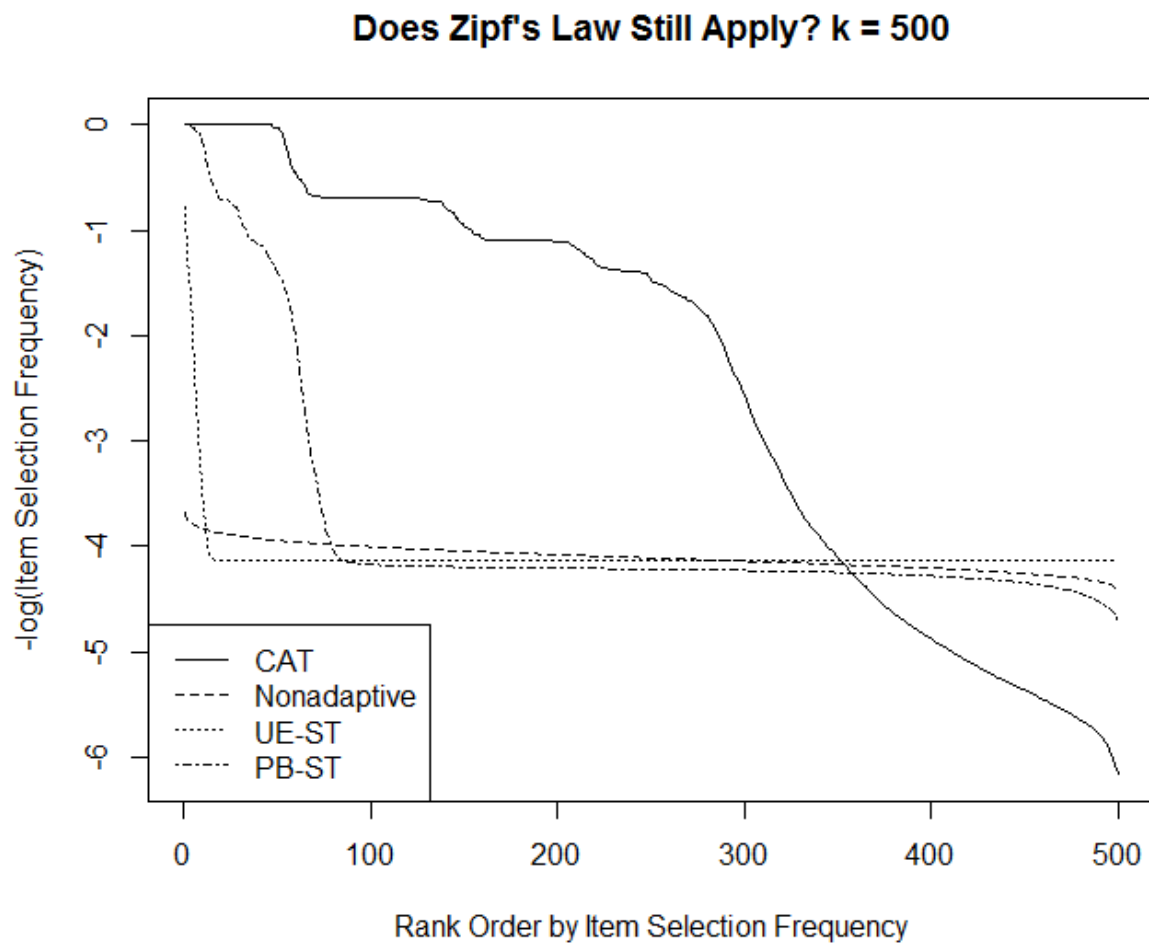


Figure 60. Rank order by item selection frequency ($k=500$)

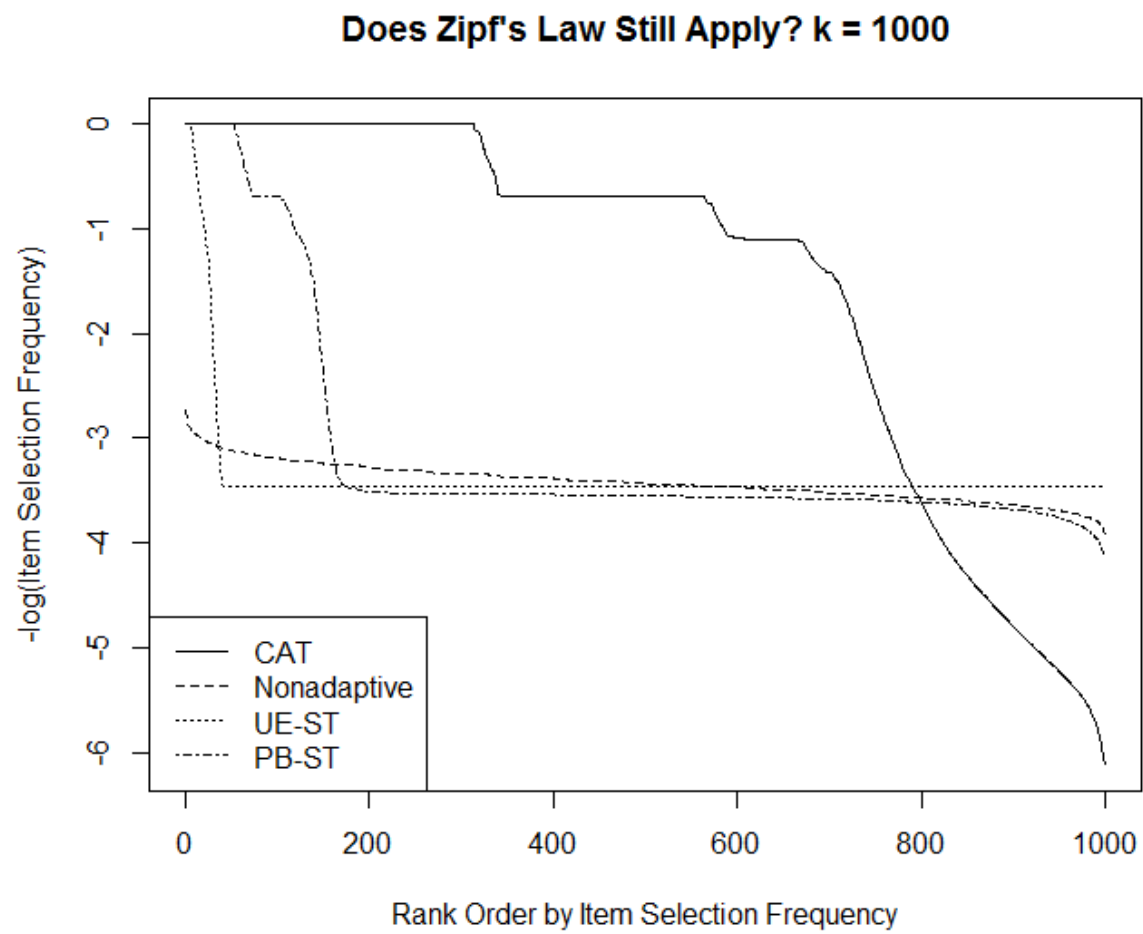


Figure 61. Rank order by item selection frequency ($k=1,000$)

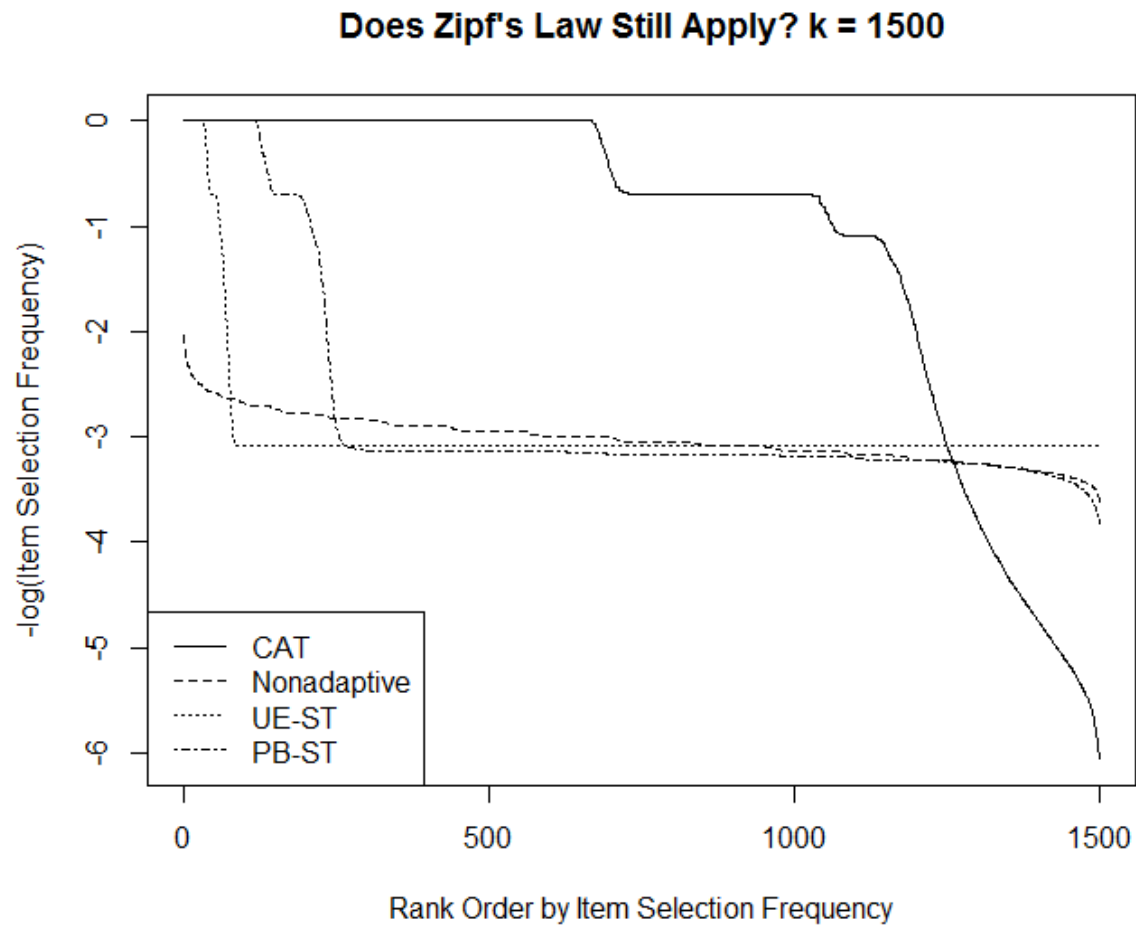


Figure 62. Rank order by item selection frequency (k=1,500)

For high-ability subjects, as an example, in a CAT one might reasonably expect others in a similar ability range to have approximately 80% items in common (if we take the modal values of the within-method overlap distributions from the results section). This is an entirely different case for PB-ST, in which (in the larger item bank conditions) even 20% overlap would be unusually high, and any items already administered are less likely to be seen in the future. Naturally, the overall effectiveness of the “remember the items you received” item pilfering strategy is determined largely by factors that are not psychometric in nature; it comes

down to the numbers, basic economics of the situation, and human cognitive ability. If pilfering is so inefficient that the money paid to take the examination allows test security to be increased more than it is compromised by a subject seeking to pilfer, then other strategies would be better avenues to explore. At a bare minimum, PB-ST achieves parity in the security versus pilfering arms race with respect to expenditure, with exponential increases in costs for gains in security than can only be negated by exponential costs on the other end. But this is all missing the point of the security issue.

There is nothing wrong with entities trying to prepare students for high-stakes examinations. There is a problem, however, when the optimal strategy is to send employees to take the exam and to remember exact test questions. An optimal strategy for performing well on the PB-ST method is not recalling the exact test questions because any examinee is unlikely to see what few questions they are able to remember. Put more practically, how many of the 1,500 items could a typical examinee remember (after the increased cost of obtaining those items)? What is easier: remembering the exact answers to even a modest 300 geometry questions, or just relearning the parts of the topic one has forgotten? It cannot be proven on a deductive basis that item pilfering is not viable, but the information available suggests that the capacity for memorization required is well outside of ordinary cognitive ability.

Related and a consistent issue in any sort of testing, but disproportionately impactful in computerized adaptive testing, is poor performance or volatile performance at the beginning of an exam. PB-ST handles this as well, as demonstrated by superior outlier control on precision, especially in relation to UE-ST. This corrective mechanism that essentially removes the damage caused by volatile early responses (as seen in the comparisons of standard error of ability by true ability for UE-ST and PB-ST) is the same exact mechanism that renders memorization

nonviable. It is not easy to misrepresent oneself with PB-ST, regardless of what direction the misrepresentation is in, and the reason for this is because misrepresentation inevitably impacts the standard error of the ability estimate.

Part 2: Limitations

Of course, there are methods that achieve other desirable outcomes -- UE-ST automatically achieves uniform item exposure, as an example. The best summary of the myriad findings from the results section is that PB-ST performs well in ability estimation, even compared to methods that are not saddled with the burden of controlling item exposure (CAT), while also performing well on exposure control, even compared to methods that specialize in it (UE-ST). For any of the featured competing methods, there is some aspect of the outcome in which PB-ST outperforms the other. One may justifiably conceptualize optimal exposure control and optimal ability estimation as mutually exclusive goals, but even the simplistic implementation of precision-based item exposure control featured through PB-ST calls this previous notion into question.

In this simplicity I refer to the fact that the “shadow tests” were of length one (rather than much longer, as they are often implemented), no projected posterior distribution of exposure frequencies was created, and no expected posterior variances were calculated to determine “need” beyond the instant an item was administered. These decisions were motivated largely by the fact that it was too computationally intensive to implement these ideas, especially when the integer programs need to be solved in a fraction of a second to be considered usable in application. PB-ST as featured here is, in fact, glorified subsetting, and this is a weakness of the study: the potential of the idea was not explored as fully as it can be.

On the other hand, it is arguably a great strength instead of a weakness that even this implementation changed the item selection process so drastically -- for evidence of this, look at every instance in the results section where UE-ST and PB-ST are compared. In PB-ST, maximum allowed exposure of an item was multiplicatively modified by the ratio of the subject's standard error of ability to an improvised metric for expected ability, whereas no such modification occurred in UE-ST. This resulted in better average precision of the ability estimate, better control of the distribution of precision of the ability estimate, trivially worsened exposure control, and the potential to scale effectively to larger item pools rather than (in the best case for UE-ST) be unaffected or the worst case negatively impacted (as seen here when sample size and test length were the same across the item bank size conditions). The useful finding here, in the sense of simplicity, is that this idea or method can be implemented even in online testing scenarios where restricted resources, server-calls for LP solutions, and other barriers might otherwise prevent a researcher from attempting a more complicated method for item exposure control. Precision-based shadow-testing, as demonstrated here, required a fraction of a second or less to perform these calculations on a modest 2.40GHz processor.

Part 3: Considerations for Implementation in Existing Testing Regimen

Executive decisions made as a necessity in the development of the simulation have to be evaluated based on the general viability of the strategies or methods used, as well as generalizable components of the simulation environment. There are a few decisions that one would be well-advised to examine critically:

1. Perhaps most critical is the mechanism by which the expected standard error is calculated. This simulation calculated it in real-time, which is not nearly as simple when subjects progress through tests at different rates, in different testing centers, etc. The

simplest way to determine this would be via simulation: take the distribution of ability observed in the population of interest and simulate a precision-based shadow test with your item bank. A sufficiently thorough simulation would provide the necessary information for use of expected standard error on at least a provisional basis. Otherwise I would generally caution against pre-determined standard error thresholds (which do not manage the item selection process in the continuous manner that PB-ST does) as well as using expected standard error calculations from other methods (discussed earlier were the ramifications if target reliability was based on estimates from a greedy CAT algorithm).

- a. It also bears mentioning that there are many different ways to calculate an expected standard error, even in the presence of usable existing data. Regardless of what precisely the method is, the important characteristic the estimate must possess is that it represents a reasonable “target reliability” that a precision-based shadow-test can strive for. An estimate that does capture a reasonable measure for central tendency in the $SE(\theta)$ distribution will produce a range of “relative reliabilities” that by proxy represent total test information for an examinee up to a specific juncture in the examination as well as whatever uncertainty was added as a result of deviations from an expected response pattern (given the estimated θ).
2. The characteristics of the item bank are important to the extent that they can make either ability estimation or exposure control difficult when incongruent to certain population parameters like ability. As a component of the pilot simulation, a uniform distribution for the b parameter was entertained as a possibility; it was not possible to both perform quality ability estimation and nearly or optimal exposure control because the difference between the distribution of difficulty and the distribution of examinee ability ($\sim N(0,1)$)

was too great. This is a classic replication of the issues one encounters when properties of the item bank are incongruent to properties of either the intended test (typically presented in the context of content constraints or similar problems) or the ability level of the population.

3. The distribution of ability in the population was simulated as standard normal, which plausibly could differ depending on the group(s) being tested. One might justifiably expect skewed or, more generally, non-normally distributed ability for self-selected, voluntary post-undergraduate examinations like the GRE, MCAT, GMAT, and others.

Part 4: Future Research Directions

When compared to CAT, we notice that PB-ST is slightly inferior with respect to ability estimation. This is counteracted by the fact that CAT's item exposure problem is so severe that Wainer (2000) suggested an entirely non-adaptive test, relinquishing the “adaptive” in “computer adaptive test”, as a plausible and potentially desirable alternative. Of course, there have been many improvements on pure CAT over the years, and this is another arguable weakness of the study; whether through limited scope or other design decisions, PB-ST was not compared to the full gamut of possible item selection strategies. The pilot study came closer to doing this, but as a pilot was limited in other ways. Admittedly, this would not be a reasonable objective for a single study, and as such this is the kind of necessary work that would be relegated to future research projects or publications.

Furthermore, arising out of the plots of overlapping items (within method) for high ability subjects, it would be useful to examine the performance of these methods where item bank size is distributed continuously – this could be particularly useful for substantive researchers

interested in determining general item bank sizes for existing or proposed item banks where overlap is deemed minimal according to whatever criteria the researcher determines.

Other potentially fruitful future research directions could be a more sophisticated implementation of precision-based exposure control, possibly using some of the concepts already discussed in this section (i.e. expected posterior variance, projected posterior frequencies of item selection, etc.), or research exploring alternative conceptualizations of what constitutes “need” in the context of computerized testing -- perhaps the ratio used here could be improved upon. Ultimately, these potential lines of research will be reliant upon continually improving computer technology for more complex formulations of these ideas as well as an ongoing need in high-stakes testing for these problems to be solved if they have not already been solved through other means.

References

- Ariel, A., van der Linden, W. J., & Veldkamp, B. P. (2006). A Strategy for Optimizing Item-Pool Management. *Journal of Educational Measurement*, 43(2), 85-96.
doi:10.1111/j.1745-3984.2006.00006.x
- Ariel, A., Veldkamp, B. P., & Breithaupt, K. (2006). Optimal Testlet Pool Assembly for Multistage Testing Designs. *Applied Psychological Measurement*, 30(3), 204-215.
doi:10.1177/0146621605284350
- Armstrong, R.D. & Jones, D. H. (1992). Polynomial algorithms for item matching. *Applied Psychological Measurement*, 16, 271-288.
- Barrada, J. R., Olea, J., & Abad, F. J. (2008). Rotating Item Banks versus Restriction of Maximum Exposure Rates in Computerized Adaptive Testing. *The Spanish Journal of Psychology*, 11(2), 618-625.
- Belov, D.I. & Armstrong, R.D. (2008). A Monte Carlo approach to the design, assembly and evaluation of multi-stage adaptive tests. *Applied Psychological Measurement*.
- Breithaupt, K., & Hare, D. (2015). Automated Test Assembly. In *Technology and Testing: Improving Educational and Psychological Measurement* (p. 128).
- Breithaupt, K., & Hare, D. R. (2007). Automated Simultaneous Assembly of Multistage Testlets for a High-Stakes Licensing Examination. *Educational and Psychological Measurement*, 67(1), 5-20. doi:10.1177/0013164406288162

- Chang, H. H., Qian, J., & Ying, Z. (2001). a-Stratified Multistage Computerized Adaptive Testing with b Blocking. *Applied Psychological Measurement*, 25(4), 333-341.
doi:10.1177/01466210122032181
- Chang, H. H., & Ying, Z. (2008). To Weight or Not to Weight? Balancing Influence of Initial Items in Adaptive Testing. *Psychometrika*, 73(3), 441-450.
- Chen, P. H., Chang, H. H., & Wu, H. (2012). Item Selection for the Development of Parallel Forms From an IRT-Based Seed Test Using a Sampling and Classification Approach. *Educational and Psychological Measurement*, 72(6), 933-953.
- Chen, S. Y., & Doong, S. H. (2008). Predicting item exposure parameters in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 61, 75-91.
- Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research*, 19, 125-136. doi:10.1007/s11136-009-9560-5
- Chuah, S. C., Drasgow, F., & Luecht, R. (2006). How Big Is Big Enough? Sample Size Requirements for CAST Item Parameter Estimation. *Applied Measurement in Education*, 19(3), 241-255. doi:10.1207/s15324818ame1903_5
- Cordova, M. J. (1997). *Optimization methods in computerized adaptive testing*. Unpublished doctoral dissertation, Rutgers University, New Brunkwick, NJ.

Edwards, M. C., Flora, D. B., & Thissen, D. (2012). Multistage Computerized Adaptive Testing With Uniform Item Exposure. *Applied Measurement in Education*.

doi:10.1080/08957347.2012.660363

Georgiadou, E., Triantafyllou, E., & Economides, A. A. (2007). A Review of Item Exposure Control Strategies for Computerized Adaptive Testing Developed from 1983 to 2005. *The Journal of Technology, Learning, and Assessment*, 5(8).

Gurobi Optimization and Inc. (2017). gurobi: Gurobi Optimizer 7.0 interface. R package version 7.0-2 URL <http://www.gurobi.com>

Hambleton, R. K., & Xing, D. (2006). Optimal and Nonoptimal Computer-Based Test Designs for Making Pass–Fail Decisions. *Applied Measurement in Education*, 19(3), 221-239.

Hau, K., & Chang, H. H. (2001). Item Selection in Computerized Adaptive Testing: Should More Discriminating Items be Used First? *Journal of Educational Measurement*, 38(3), 249-266. doi:10.1111/j.1745-3984.2001.tb01126.x

Kalinowski, K. E., Natesan, P., & Henson, R. K. (2014). Stratified Item Selection and Exposure Control in Unidimensional Adaptive Testing in the Presence of Two-Dimensional Data. *Applied Psychological Measurement*, 38(7), 563-576.

Kim, J., Chung, H., & Dodd, B. G. (2012). Panel Design Variations in the Multistage Test Using the Mixed-Format Tests. *Educational and Psychological Measurement*.

doi:10.1177/0013164411428977

- Kim, S., & Moses, T. (2014). *An Investigation of the Impact of Misrouting Under Two-Stage Multistage Testing: A Simulation Study*. ETS Research Report Series 2014.1.
- Kim, S., Moses, T., & Yoo, H. (2015). A Comparison of IRT Proficiency Estimation Methods Under Adaptive Multistage Testing. *Journal of Educational Measurement*, 52(1), 70-79.
- Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2(4), 359–375.
- Lissitz, R. W., & Jiao, H. (2012). *Computers and their impact on state assessments*. Charlotte, NC: Information Age Publishing, Inc.
- Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Magis, D., & Raiche, G. (2012). Random Generation of Response Patterns under Computerized Adaptive Testing with the R Package catR. *Journal of Statistical Software*, 48(8), 1-31.
URL <http://www.jstatsoft.org/v48/i08/>.
- McBride, J. R. & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing* (pp.223-226). New York, Academic Press.
- R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>
- Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, 26(2), 144-163.

- Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glass (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 163-182). Netherlands: Kluwer Academic Publishers.
- van der Linden, W. J. (2005). A Comparison of Item-Selection Methods for Adaptive Tests with Content Constraints. *Journal of Educational Measurement*, 42(3), 283-302.
- van der Linden, W. J. (2008). Using Response Times for Item Selection in Adaptive Testing. *Journal of Educational and Behavioral Statistics*, 33(1), 5-20.
doi:10.3102/1076998607302626
- van der Linden, W. J., Ariel, A., & Veldkamp, B. P. (2006). Assembling a Computerized Adaptive Testing Item Pool as a Set of Linear Tests. *Journal of Educational and Behavioral Statistics*. doi:10.3102/10769986031001081
- van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting Differential Speededness in Multistage Testing. *Journal of Educational Measurement*, 44(2), 117-130. doi:10.1111/j.1745-3984.2007.00030.x
- van der Linden, W. J., & Glas, C. A. (2010). *Elements of adaptive testing*. New York: Springer.
- van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22, 259-270.
- Wainer, H. (2000). Rescuing Computerized Testing by Breaking Zipf's Law. *Journal of Educational and Behavioral Statistics*, 25(2), 203-224. doi:10.3102/10769986025002203

- Wang, C., Chang, H. H., & Boughton, K. A. (2012). Deriving Stopping Rules for Multidimensional Computerized Adaptive Testing. *Applied Psychological Measurement*, 37(2), 99-122. doi:10.1177/0146621612463422
- Wang, C., Zheng, Y., & Chang, H. H. (2014). Does Standard Deviation Matter? Using "Standard Deviation" to Quantify Security of Multistage Testing. *Psychometrika*, 79(1), 154-174. doi:10.1007/S11336-013-9356-Y
- Wise, S. L., & Kingsbury, G. G. (2000). Practical Issues in Developing and Maintaining a Computerized Adaptive Testing Program. *Psicologica*, 21, 135-155.
- Yan, D., von Davier, A. A., & Lewis, C. (Eds.). (2014). *Computerized multistage testing: Theory and applications*.
- Yi, Q., Zhang, J., & Chang, H. H. (2008). Severity of Organized Item Theft in Computerized Adaptive Testing: A Simulation Study. *Applied Psychological Measurement*, 32(7), 543-558. doi:10.1177/0146621607311336
- Zenisky, A. L. (2004). *Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment* (Unpublished doctoral dissertation). University of Massachusetts, Amherst.
- Zenisky, A. L., Hambleton, R. K., & Sired, S. G. (2002). Identification and Evaluation of Local Item Dependencies in the Medical College Admissions Test. *Journal of Educational Measurement*. doi:10.1111/j.1745-3984.2002.tb01144.x

Zheng, Y., & Chang, H. H. (2015). On-the-Fly Assembled Multistage Adaptive Testing. *Applied*

Psychological Measurement, 39(2), 104-118. doi:10.1177/0146621614544519

Zhou, X., & Reckase, M. D. (2014). Optimal item pool design for computerized adaptive tests

with polytomous items using GPCM. *Psychological Test and Assessment Modeling*,

56(3), 255-274.